GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

InfiniBand Fabric Health: Counters and Fault Isolation

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-13 · infiniband fabric health · infiniband troubleshooting · fabric health check · topology discovery · port counters · ibdiagnet · gpu clusters · fault isolation · rdma

Original article: https://gpusmith.com/articles/en/infiniband-fabric-health-fault-isolation

In this article

1. Executive Summary
2. Introduction and Background
3. What Fabric Health Means
4. A Read-Only InfiniBand Fabric Health Baseline
5. InfiniBand Topology Discovery and Comparison
6. Port Counters: Meanings, Deltas and Limits
7. A Fault-Isolation Runbook for GPU Clusters
8. ibdiagnet and the Escalation Bundle
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

InfiniBand fabric health is an evidence question: does the observed fabric match its intended connectivity and link capabilities, and do new counter events explain the affected workload? **ibnetdiscover** provides topology identifiers and connectivity, **iblinkinfo** reports links, **ibqueryerrors** applies reporting thresholds, and **perfquery** queries individual port counters. Together they support a reproducible starting point for operating high performance computing (HPC) and graphics processing unit (GPU) clusters. (man7.org) (manpages.debian.org) (manpages.ubuntu.com) (manpages.ubuntu.com) The recommended sequence is to preserve topology and counters, measure changes, correlate affected endpoints, and then test a specific fault hypothesis.

Two numerical details prevent common misinterpretations. The default **ibqueryerrors threshold is 0**, which is a reporting choice, not a hardware health limit. (manpages.ubuntu.com) Standard performance data counters represent **octets divided by 4**, so byte-rate calculations require that conversion. (man7.org) NVIDIA's documented **ibdiagnet 2.19** sampling interval defaults to **1 second**; setting its interval to **0** suppresses the second sample. That convenience does not establish a suitable observation window for a particular collective workload. (networking-docs.nvidia.com) Report absolute counts and interval deltas separately, and avoid inferring a continuous rate across an unexplained reset or identity change.

A link hypothesis becomes more persuasive when physical-lane errors or recovery events increase on an identified connection during the affected job. Successful recovery and link-down events have different meanings, while transmit discards can have several causes. (docs.oracle.com) (docs.nvidia.com) (docs.nvidia.com) (docs.nvidia.com) Conversely, quiet counters should prompt checks of device selection, locked memory and GPU proximity to the network adapter. Unified Communication X (UCX) supports explicit device selection, and Cambridge's published GPU-cluster guidance demonstrates why the closest network interface matters. (openucx.readthedocs.io) (docs.hpc.cam.ac.uk) Topology-aware job placement also deserves comparison with actual fabric connectivity. (slurm.schedmd.com)

This report provides a read-only-first baseline, a command-to-question table, a counter glossary, a topology-diff procedure and an incident worksheet. It treats replacement and counter clearing as controlled interventions after evidence capture. **No universal error-rate threshold or collective bandwidth target is proposed.** Operators should preserve the before-and-after evidence, validate command options against the installed stack, and escalate with port identity, peer identity, workload timing and a record of every intervention.

Introduction and Background

InfiniBand is a **channel-based switched fabric** for server and storage connectivity, according to the InfiniBand Trade Association (IBTA). Its subnet manager provides routing functionality. (infinibandta.org) (infinibandta.org) A health check therefore needs more than a host's report that a connection exists: it needs a view of the participating devices, their connections and the operating evidence relevant to the workload.

The operational question is usually narrower than whether InfiniBand is a suitable fabric. A previously acceptable collective slows down, selected jobs show uneven progress, or performance changes after maintenance. The operator must decide which hypothesis to test first: a physical link, an unexpected connection, a shared path under pressure, or a host-side configuration. This report proposes an evidence sequence for that decision. Its recommendations are procedures, not reported benchmark results.

Every InfiniBand subnet needs a running **subnet manager (SM)**. SUSE documents that requirement, while Red Hat explicitly includes a direct connection between machines without a switch. (documentation.suse.com) (docs.redhat.com) Record the active manager and its state before interpreting a partial discovery. Oracle documents **sminfo** for querying manager priority and state, and describes manager responsibilities including port local identifier (LID) assignment and forwarding-table configuration. (docs.oracle.com) (docs.oracle.com)

The dated reference point is **October 13, 2026**. Documentation reviewed here includes distribution manuals and versioned vendor references; these do not prove that any particular package is installed in the reader's cluster. An operating procedure should record the actual tool, driver, firmware and management-stack versions before using the examples below.

GPU Smith's published engineering scope includes operational telemetry and procedures, and its first-party documentation emphasizes port maps and as-built records. That provides an adjacent engineering perspective: retain an inventory that lets an operator connect a diagnostic identifier to physical equipment. (gpusmith.com) (gpusmith.com) The article's technical conclusions rely on tool maintainers, operating-system vendors, standards guidance and institutional documentation rather than claims of consultancy deployments.

What Fabric Health Means

Separate identity, connectivity, counters and workload

The diagnostic baseline should answer separate questions instead of collapsing them into one green status. **Globally unique identifiers (GUIDs)**, LIDs, node types and port numbers appear in topology discovery output. (man7.org) Linux separately exposes logical port state and physical state; its logical states include DOWN, INIT, ARMED and ACTIVE. (www.kernel.org) These are related observations, but each deserves its own worksheet field.

Use the following categories as an operational taxonomy:

- **Identity:** Which host channel adapter (HCA), local port and subnet supplied the observations? Explicit selection matters because the diagnostics library otherwise searches for an ACTIVE port. (manpages.debian.org)
- **Connectivity:** Does each recorded endpoint connect to the intended peer, including the intended switch port? Discovery and connection comparison supply that evidence. (man7.org)
- **Link capability:** Does active width and speed match the installed design? Discovery includes those attributes, and Linux describes active data rate as width multiplied by speed. (man7.org) (www.kernel.org)
- **Counter behavior:** Which events and traffic counters changed during the observation interval? Performance-management queries supply the observations; the procedure must supply consistent timing. (manpages.ubuntu.com)
- **Workload relationship:** Which allocated nodes, rails and placements share the relevant connection? Slurm supports topology-aware allocation, making scheduler placement a useful companion record. (slurm.schedmd.com)

The recommended health record is therefore a collection of observations, not a single score. A correct peer with a changed link capability deserves attention even before an error threshold fires. A stable topology with growing wait counters deserves another line of inquiry. A job that selects an unintended host interface needs a host test rather than an immediate cable replacement.

Preserve the measurement context

Read-only does not mean permissionless. Diagnostic access may require a management key (MKey), and some diagnostic operations are restricted from virtual machines. (manpages.debian.org) NVIDIA's current command reference also exposes a performance-management key, **pm_key**. (networking-docs.nvidia.com) Treat authorization failures as missing observations, not as proof that the affected ports have no errors.

Define the scope before collection: subnet, rail, collector host, local HCA port, job interval and intended topology revision. If the management team cannot explain which subnet a command scanned, its clean output should remain provisional. Recommendation: retain standard output, standard error, return status and the exact command together. That makes a partial discovery or unsupported attribute distinguishable from an actual zero result.

A Read-Only InfiniBand Fabric Health Baseline

Validate the installed tool set

The maintained **infiniband-diags** collection is intended for fabric configuration, debugging and maintenance. (manpages.debian.org) Package names should come from the target distribution rather than a generic installation recipe. Red Hat Enterprise Linux (RHEL) documentation distinguishes **rdma-core** from verification packages **libibverbs-utils** and **infiniband-diags**. (docs.redhat.com) (docs.redhat.com) Confirm installed package versions locally rather than treating an online manual as proof of the node's software.

Before the first scan, capture these baseline items:

- **Software:** Distribution release, package versions, tool help output and firmware identifiers. Retain the local evidence rather than inferring versions from an online manual.
- **Collector:** Hostname, HCA name, selected local port and rail. Use explicit adapter and port selection where the installed command supports it. (manpages.debian.org)
- **Manager:** SM identity, priority and state, plus the management team's record of recent failover or configuration changes. (docs.oracle.com)
- **Inventory:** Physical host and switch labels joined to GUIDs. Node-name maps provide a documented GUID-based lookup mechanism. (manpages.debian.org)
- **Design:** Intended endpoint pairs and expected width/speed for each class of connection. Active attributes belong in the observed record. (man7.org)
- **Jobs:** Affected allocations, rank placement, start/end times and known-good comparisons. Topology-aware scheduling makes this context relevant. (slurm.schedmd.com)
- **Collection result:** Raw outputs, timestamps, exit status, permission errors and unreachable devices. Keep incomplete observations visibly incomplete.

Table 1 maps the baseline commands to specific questions. Options reflect the opened manuals as of October 13, 2026; operators should confirm them locally before execution.

Command	Question answered	Evidence-preserving use
ibnetdiscover	What nodes and port connections are visible?	Save the raw topology and a discovery cache when supported; keep both for comparison. (man7.org) (manpages.debian.org)
iblinkinfo	What link information is reported across the fabric?	Save normal output and a one-link-per-line view using `--line`; preserve warnings. (manpages.debian.org) (manpages.debian.org)
ibqueryerrors	Which port counters exceed the chosen reporting thresholds?	Save the selected threshold file and request details when supported; omit clearing options. (manpages.ubuntu.com) (man7.org)
perfquery	What are the counters for a specific identified port?	Query basic and supported extended attributes; retain addressing mode and queried port. (manpages.ubuntu.com) (manpages.ubuntu.com)

The commands answer complementary questions. A filtered error report does not replace topology discovery, and a discovery output does not establish that a job used those connections. Use the table as a collection checklist, then join the results by identity.

InfiniBand Topology Discovery and Comparison

Capture first, compare second

The following shell template is an operational example, not a remote action executed by this report. Replace the adapter, local port and destination directory with approved local values. The manuals document local adapter selection; confirm the installed syntax before running it. (manpages.debian.org)

```
mkdir -p /approved/incident-directory
date -u > /approved/incident-directory/collection-start.txt
ibnetdiscover -C mlx5_0 -P 1 \
  > /approved/incident-directory/topology.txt \
  2> /approved/incident-directory/topology.stderr
iblinkinfo -C mlx5_0 -P 1 --line \
  > /approved/incident-directory/links.txt \
  2> /approved/incident-directory/links.stderr
ibqueryerrors -C mlx5_0 -P 1 --details \
  > /approved/incident-directory/errors.txt \
  2> /approved/incident-directory/errors.stderr
date -u > /approved/incident-directory/collection-end.txt
```

Retain each command's return status in the production wrapper. Replace the capitalized arguments below with verified local values. Once a target has been identified, a non-resetting per-port query can be made with `perfquery -C HCA_NAME -P LOCAL_PORT -G TARGET_PORT_GUID TARGET_PORT`; add `-x` only for a supported extended query. **-G means Port GUID addressing**, so do not substitute a node GUID merely because its label looks familiar. (manpages.ubuntu.com) (manpages.ubuntu.com)

A topology-diff procedure

Use a stored intended baseline and a current snapshot. Upstream ibnetdiscover supports comparison of switches, channel adapters, routers and connections; iblinkinfo has separate diff checks including port state and LIDs. (man7.org) (manpages.debian.org) Recommendation:

- **Preserve originals:** Retain the prior snapshot, new snapshot and their collection metadata before any normalization.
- **Join by identifiers:** Compare GUID and port pairs, then resolve physical labels through the inventory map. (manpages.debian.org)
- **Classify changes:** Separate node membership and connection changes from addressing or descriptive changes. The discovery tool exposes selectable comparison categories. (manpages.debian.org)
- **Compare capability:** Review width and speed separately from peer identity. (man7.org)
- **Compare state:** Use the link tool's documented state comparison rather than copying its diff keys blindly into the discovery tool. (manpages.debian.org)
- **Explain every difference:** Attach a maintenance record or investigation hypothesis; retain unexplained differences for isolation.

An ordinary text diff is useful for review, but do not let reordered output become a supposed cable move. The engineering objective is a changed endpoint relationship or changed operating property, supported by the preserved raw records.

Port Counters: Meanings, Deltas and Limits

A glossary for operational decisions

Counters require their documented meaning before they acquire an incident interpretation. **Symbol errors** describe physical-lane detections, whereas **link error recovery** counts successful recovery completions. (docs.oracle.com) (docs.nvidia.com) Successful recovery is evidence of an event; it is not equivalent to an unchanged link.

Table 2 groups the key observations. Definitions reflect the opened Linux, Oracle and NVIDIA documentation as of October 13, 2026; escalation choices are recommendations rather than vendor health limits.

Counter or observation	Documented meaning	Recommended interpretation
SymbolErrorCounter	Symbol errors detected on physical lanes. (docs.oracle.com)	Examine interval growth, peer evidence and active capability before selecting a physical-link test.
LinkErrorRecoveryCounter	Successful completion of link-error recovery by the training state machine. (docs.nvidia.com)	Correlate new recoveries with workload timing and link observations.
LinkDownedCounter	Recovery failed and the link was downed. (docs.nvidia.com)	Preserve both endpoint records and any contemporaneous state changes.
PortXmitDiscards	NVIDIA lists causes including an inactive output port, excess over the neighbor maximum transmission unit (NeighborMTU), and switch-lifetime limits. (docs.nvidia.com)	Ask for discard details; avoid assigning every discard to congestion.
PortRcvErrors and discard details	The error-query tool can retrieve receive-error and transmit-discard detail counters. (man7.org)	Use detail attributes to narrow the hypothesis when supported.
VL15Dropped	Oracle documents incoming virtual-lane 15 packets dropped for lack of buffers. (docs.oracle.com)	Treat it as a virtual-lane-specific observation requiring context, rather than a direct application throughput measurement.
PortXmitData / PortRcvData	Data units are four octets; Linux's corresponding sysfs data counters are 64-bit. (docs.nvidia.com) (www.kernel.org)	Convert the correct attribute's delta to bytes; retain its provenance and interval.
PortXmitWait	Ticks with data available but no transmission because credits or arbitration were insufficient. (www.kernel.org)	Examine pressure and workload context; establish hardware-specific tick meaning before converting to time.
Logical / physical state	Linux exposes these separately. (www.kernel.org)	Record both alongside expected and observed width/speed.

The glossary separates event evidence from activity evidence. High transmit volume is not itself an error. A wait observation does not identify a damaged cable, and a recovery event does not identify which replaceable component is responsible. The next useful question is what changed, where, and while which workload was running.

Reporting thresholds are not acceptance limits

The default error-query threshold is **zero**. An operator can choose threshold values through a file, but that changes what the tool prints. (manpages.ubuntu.com) Recommendation: store that file with the incident and document exclusions. A quiet filtered report means no reported counter exceeded the chosen filter under the scan's conditions; it should not become an unconditional health certificate.

Obtain **baseline and follow-up samples** for the same port and attribute. Preserve absolute values so that another engineer can distinguish a long-lived historical count from interval growth. Use the workload's real timing to choose the observation interval. A short sample can answer a short-event question; a representative collective comparison needs representative workload context.

Counter handling must remain attribute-specific. Extended counters are optional, and perfquery documents explicit reset operations. (man7.org) (man7.org) Linux's documented 64-bit sysfs data counters do not establish the width or rollover behavior of every hardware error counter. (www.kernel.org) Record counter width and reset/saturation semantics from the deployed hardware documentation when available. If the follow-up value falls, changes identity, or becomes unsupported, classify the interval as discontinuous until explained. **Do not apply a universal modulo correction.**

A Fault-Isolation Runbook for GPU Clusters

Correlate before replacing

The recommended isolation sequence connects fabric evidence to affected allocations. Slurm's topology support can inform resource selection, but its optional topology LinkSpeed field has arbitrary units and is documented as unused. It should not be treated as a measured link-speed sensor. (slurm.schedmd.com) (slurm.schedmd.com) Preserve scheduler configuration and actual port observations separately.

A collective slowdown should lead to a reproducible comparison with controlled node placement, message sizes and software settings. Lawrence Livermore National Laboratory (LLNL) explicitly labels its published measurements as taken on live, heavily loaded systems using a specified Message Passing Interface (MPI) test. That illustrates why test conditions belong beside a bandwidth figure. (hpc.llnl.gov) The procedure below proposes controls; it reports no new performance results.

- **Confirm scope:** Identify affected jobs, selected HCAs and rails; establish whether observations came from the intended subnet. (manpages.debian.org)
- **Check topology:** Review unexplained endpoint or capability differences before narrowing to an individual counter. (man7.org) (man7.org)
- **Sample both ends:** Preserve target and peer counters with query times, including errors and relevant traffic attributes. (man7.org) (man7.org)
- **Test a physical hypothesis:** If evidence points to a connection, resolve diagnostic GUID/port to the physical connector before touching hardware. Oracle explicitly distinguishes those identifiers from connector labels. (docs.oracle.com)

- **Control the intervention:** Under the site's change process, change one candidate component or path at a time. Oracle's hardware-specific guidance uses a known-good cable to isolate a suspect connection. (docs.oracle.com)
- **Test a pressure hypothesis:** Where wait or discard growth accompanies a shared workload, compare job placement and load while retaining the same collection method. The wait definition includes both credits and arbitration. (www.kernel.org)
- **Test a host hypothesis:** Compare the affected host's device visibility, memory limits and runtime selection with a known-good host. Open MPI's troubleshooting guidance includes port checks and locked-buffer requirements. (docs.open-mpi.org)
- **Reobserve:** Repeat the same workload and collection method after the intervention. Decide whether the symptom and the relevant new events recur before accepting the result.

Recommendation: after preserving evidence, drain or move affected workloads under the site's change process. Check labels and seating, then test a compatible known-good cable or optic assembly, a known-good switch port, an HCA port, and finally the host, as the evidence permits. Restore the intended topology between tests and change one variable at a time. Oracle's cable-isolation procedure supports the known-good substitution principle; this ordering is proposed operational guidance. (docs.oracle.com)

A component replacement is an experiment. If the observations follow a changed cable or port, preserve that relationship; if they remain with the host, move the host hypothesis up the list. Do not describe this as a proven root cause until the evidence distinguishes competing explanations.

Host-side checks with a clean-looking fabric

For Open MPI builds with UCX support, the opened documentation describes UCX selection for InfiniBand, and its network-support page identifies the UCX point-to-point messaging layer (PML). (docs.open-mpi.org) (docs.open-mpi.org) Record what the installed application actually selected rather than assuming that the presence of an HCA proves its use.

UCX supports restricting devices with **UCX_NET_DEVICES**, while **ucx_info** provides device and configuration introspection. (openucx.readthedocs.io) (openucx.readthedocs.io) Preserve the unrestricted configuration before a controlled single-rail test. Explicit transport restrictions also need care: UCX warns that GPU-memory applications must retain relevant GPU transports. (openucx.readthedocs.io) An intervention that changes device selection and GPU-memory handling simultaneously is difficult to interpret.

Non-uniform memory access (NUMA) locality is another host hypothesis. UCX discusses inefficient GPU zero-copy remote direct memory access (RDMA) when the GPU and adapter occupy different NUMA nodes. (openucx.readthedocs.io) Cambridge's Wilkes3 guidance recommends the closest network interface card (NIC) and closest central processing unit (CPU) domain; its published layout has a NIC shared by a pair of GPUs. This is a site-specific illustration, not a wiring template for other systems. (docs.hpc.cam.ac.uk) (docs.hpc.cam.ac.uk)

Finally, RDMA requires pinned memory, and containers need access to corresponding RDMA device nodes. (docs.redhat.com) (openucx.readthedocs.io) Record **ulimit -l**, device exposure and launch configuration. Quiet fabric counters should keep these possibilities open rather than end the investigation.

ibdiagnet and the Escalation Bundle

Add vendor diagnostics without losing the baseline

ibdiagnet can supplement the baseline with vendor-documented diagnostic collection. The opened **2.19** command reference is dated **June 25, 2026**. It documents **--pc as a fabric port-counter reset**, including standard and extended counter sets. (networking-docs.nvidia.com) That option belongs after preservation of evidence, not in the first health snapshot.

The same reference documents **--pm_pause_time** for the interval between samples, with a default of **1 second** and a **0** setting that omits the second sample. (networking-docs.nvidia.com) Choose an interval appropriate to the question instead of treating the default as a vendor performance criterion. Keep the installed version and exact invocation beside any generated output.

For physical-layer investigation, NVIDIA documents **--get_phy_info** to query physical-layer (PHY) information, including bit error rate (BER) counters. Its separate **--reset_phy_info** operation queries and clears that information. (networking-docs.nvidia.com) (networking-docs.nvidia.com) The documented collection outputs include **ibdiagnet2.db_csv** and **ibdiagnet2.net_dump_ext**. (networking-docs.nvidia.com) Save the raw artifacts and use deployed-device documentation before comparing lane-specific observations.

The reset distinction applies across tool families. **ibqueryerrors --clear-errors / --clear-counts** can clear counters whether or not they were printed, while perfquery can reset after reading or reset without reading. (man7.org) (man7.org) Ubuntu's perfquery manual also illustrates selective reset masks. (manpages.ubuntu.com) Do not copy a reset mask into a generic first-response command, and do not assume similarly named short options mean the same thing in different tools.

A reusable incident worksheet

Table 3 specifies a proposed worksheet, not a vendor schema. It joins diagnostic identifiers, physical inventory and workload context; those categories follow the discovery, connector-mapping and scheduling evidence. (man7.org) (docs.oracle.com) (slurm.schedmd.com)

Field group	Fields to retain	Decision supported
Identity	Subnet/rail; collector; local HCA/port; node GUID; port GUID; LID	Was the same endpoint measured through the intended management path?
Connection	Switch label/port; peer GUID/port; cable or optic inventory ID	Can the suspect observation be mapped to equipment without ambiguity?
Capability	Expected and active width/speed; logical and physical state	Is the link operating as the design expects?
Counter sample	Counter name; attribute/source; baseline; follow-up; timestamps; delta; known reset semantics	Is interval growth valid and relevant?
Workload	Job ID; nodes/ranks; selected devices; placement; message/workload parameters	Does the observation coincide with the affected communication?
Change and action	Topology revision; maintenance change; proposed test; intervention time; after-test evidence	Did a controlled change alter both symptom and evidence?

The worksheet should make unknowns explicit. An empty peer field, uncertain reset history or missing sample is a gap to resolve, not a zero. Avoid merging observations from different rails merely because their hostnames match.

The recommended escalation bundle contains:

- **Time record:** Incident interval, collector clock reference, scan start/end and intervention times.
- **Version record:** Operating system, installed packages, firmware, HCA identity, tool help and runtime settings.
- **Manager record:** SM identity/state and relevant configuration or change history. (docs.oracle.com)
- **Topology record:** Intended map, raw discovery, stored cache and classified differences. (manpages.debian.org)
- **Port record:** Both endpoint identities, physical labels, raw samples and selected threshold file. (docs.oracle.com) (manpages.ubuntu.com)
- **Workload record:** Allocation, runtime device choice, reproducible comparison and environmental constraints.
- **Action record:** Exactly what changed, which counters were cleared, by whom or which automation, and whether new observations recurred.

Data Analysis and Evidence

Convert counters into a defensible local rate

The quantitative foundation is the documented unit, not a claimed market-wide benchmark. In the performance attributes described by perfquery, data counters measure octets divided by **4**. Linux also documents this unit for its sysfs data counters. (man7.org) (www.kernel.org) For a continuous interval measured in seconds, the proposed calculation is:

```
delta_units = follow_up_units - baseline_units
delta_bytes = 4 * delta_units
average_bytes_per_second = delta_bytes / elapsed_seconds
average_Gbit_per_second = 8 * delta_bytes / elapsed_seconds / 1000000000
```

Hypothetical Example: Assume a data counter moves from **2,000,000,000** to **2,600,000,000 units** over **20 seconds**, with the same endpoint and no reset or discontinuity. The author calculation yields **2,400,000,000 bytes**, **120,000,000 bytes/second**, or **0.96 Gbit/second** in decimal units. These inputs and results are illustrative arithmetic, not measured hardware performance or an acceptance limit. The conversion factor comes from the documented counter units. (man7.org)

Traffic measured at a port should not be relabeled as collective application bandwidth. A controlled workload comparison should state what its measurement represents. LLNL describes a bandwidth test using a task on each of a pair of nodes; that method is useful context, but its historical platform results are not proposed as current GPU-cluster targets. (hpc.llnl.gov)

Distinguish an event delta from a health threshold

Hypothetical Example: Assume a symbol-error counter rises from **120** to **132** during the same **20-second** observation. The author calculation is a delta of **12 events**, or **0.6 events/second**. The counter's physical-lane meaning is documented, but these counts are invented example inputs. Neither the resulting rate nor a zero-default query filter establishes the deployed hardware's acceptance limit. (docs.oracle.com) (manpages.ubuntu.com)

Useful local comparisons pair the affected interval with a known-good interval under stated conditions. Preserve node placement, link capability, traffic volume, counter source and collection timing. If those conditions differ, present the difference rather than hide it inside a health score. This is a proposed comparison method, not an assertion that historical baselines guarantee future performance.

For memory-related isolation, UCX's container guidance recommends at least **128 MB** of locked-memory allowance for its resources and buffers, while warning that a larger application working set can need more. That is a configuration recommendation, not an InfiniBand fabric limit. (openucx.readthedocs.io) Open MPI likewise explains the locked-buffer requirement. (docs.open-mpi.org) Record the actual application allowance instead of copying the example into every host configuration.

The evidence supports units, command behavior and a reproducible investigation method. It does not support a universal error budget, a hardware-independent wait-tick conversion or a promised collective throughput. Where saturation, rollover or reset semantics remain undocumented for the deployed attribute, preserve the raw observations and escalate that uncertainty explicitly.

Implications and Future Directions

Fabric health monitoring becomes more useful when it preserves context that a one-time scan lacks. The recommendation is to retain an intended topology revision, named inventory, manager identity, port observations and workload metadata together. Diagnostic node-name maps already support GUID-based naming, and Slurm already exposes topology-aware scheduling concepts. (manpages.debian.org) (slurm.schedmd.com) The engineering work is to keep their records consistent with the actual installation.

Counter automation should also preserve provenance. Linux's 64-bit sysfs data definition and perfquery's optional extended attributes illustrate why collectors should identify the attribute they read instead of publishing an unexplained number with a familiar counter name. (www.kernel.org) (man7.org) A future dashboard should expose interval boundaries and discontinuities rather than silently turn every falling value into a positive rate.

Changes to hardware, firmware or management software should trigger a review of the collection profile. PHY queries and PHY resets are separate documented operations, which makes reset policy an essential part of automation review. (networking-docs.nvidia.com) (networking-docs.nvidia.com) Proposed production practice: keep querying and clearing in separate procedures, with the latter recording its scope and time. This protects subsequent interpretation without prescribing a universal maintenance interval.

Host and fabric records should remain connected. Red Hat documents separate user access to RDMA verbs, and Open MPI documents its network-selection behavior; neither substitutes for an observed application path. (docs.redhat.com) (docs.open-mpi.org) Cambridge's placement guidance shows why the GPU-to-adapter

relationship is useful incident context even when the question begins at a switch port. (docs.hpc.cam.ac.uk)

GPU Smith's first-party emphasis on as-built port maps is consistent with that documentation requirement. (gpusmith.com) For an adjacent engineering advisor, the relevant contribution is a reproducible evidence package and explicit assumptions. Health decisions should remain tied to the installed equipment and workload, with uncertain counter semantics or missing measurements stated openly.

Frequently Asked Questions (FAQs)

What is the fastest useful InfiniBand fabric health check?

Start by confirming scope and manager state, preserving discovery and link output, then collecting error and target-port counters without clearing. The tools provide complementary observations; the recommendation is to join them before taking a disruptive action. (docs.oracle.com) (man7.org) (manpages.debian.org) (man7.org)

Does a nonzero symbol-error counter prove a bad cable?

It establishes recorded physical-lane errors, not a unique replaceable component. Measure new events, identify both ends and map their physical connectors before designing an isolation test. (docs.oracle.com) (docs.oracle.com) A known-good cable comparison is one documented hardware-specific technique. (docs.oracle.com)

Why can GPU collectives slow down when error reports look clean?

The report may be filtered, or the application may have a host-side constraint. Check thresholds, runtime device selection, memory registration requirements and GPU-adapter locality. (manpages.ubuntu.com) (openucx.readthedocs.io) (docs.open-mpi.org) (openucx.readthedocs.io) Clean-looking counters should be one observation in that investigation.

When should counters be cleared?

After saving the evidence and agreeing the scope of a controlled recurrence test. The tools document resets, including clear-after-read behavior that can affect counters absent from printed output. Record the clear time and method so the next interval has an explainable baseline. (man7.org) (man7.org)

Conclusion

A useful InfiniBand health assessment ends with a testable explanation and a preserved record. The recommended baseline begins with the collector, subnet manager and intended fabric, then captures topology, operating capability and counter values before any intervention. The operator should be able to connect each observation to an endpoint, its peer, a physical label and an affected workload.

The main distinction is between recorded history and new events under known conditions. Reporting thresholds select output; they do not create hardware acceptance limits. Traffic conversion requires documented units, while event counters require documented meanings. Unexplained resets, identity changes

or unsupported attributes should remain visible as measurement gaps.

Fault isolation should progress through a specific hypothesis at a time. A physical-link test needs physical mapping and peer evidence. A pressure hypothesis needs workload and path context. A host hypothesis needs actual device selection, memory constraints and placement information. Each intervention should be followed by comparable observations rather than accepted merely because a command completed.

The durable outcome is the escalation bundle: raw topology, meaningful differences, target and peer samples, workload timing, version information and an action record. That bundle supports the next engineering decision and makes recurrence easier to assess. It also keeps the limits of the evidence explicit, allowing an operating team to investigate without replacing uncertainty with an unsupported health score.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://man7.org/linux/man-pages/man8/ibnetdiscover.8.html
2. https://manpages.debian.org/trixie/infiniband-diags/iblinkinfo.8.en.html
3. https://manpages.ubuntu.com/manpages/noble/man8/ibqueryerrors.8.html
4. https://manpages.ubuntu.com/manpages/noble/en/man8/perfquery.8.html
5. https://man7.org/linux/man-pages/man8/perfquery.8.html
6. https://networking-docs.nvidia.com/ibdiagnetutilityum/219/basic-commands
7. https://docs.oracle.com/cd/E76425_01/html/E36257/z40000061891278.html
8. https://docs.nvidia.com/networking/display/ufmenterpriseumv6191/appendix+-+supported+port+counters+and+events
9. https://gpusmith.com/articles/en/gpu-cpu-nic-affinity-rank-binding
10. https://openucx.readthedocs.io/en/master/faq.html
11. https://docs.hpc.cam.ac.uk/user-guide/a100.html
12. https://slurm.schedmd.com/topology.html
13. https://infinibandta.org/ibta-specification/
14. https://gpusmith.com/articles/en/infiniband-vs-spectrum-x-vs-roce-vs-ethernet-ai-clusters
15. https://documentation.suse.com/sles/16.0/html/SLES-subnet-manager/index.html
16. https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/10/html/configuring_and_managing_networking/configuring-infiniband-and-rdma-networks
17. https://docs.oracle.com/cd/E18476_01/doc.220/e18478/GUID-9FF8B5B0-3481-4B73-89D3-108CBD7EB989.htm
18. https://gpusmith.com
19. https://www.kernel.org/doc/html/latest/admin-guide/abi-stable.html
20. https://manpages.debian.org/trixie/infiniband-diags/infiniband-diags.8.en.html
21. https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/8/html-single/configuring_infiniband_and_rdma_networks/index
22. https://manpages.debian.org/trixie/infiniband-diags/ibnetdiscover.8.en.html
23. https://man7.org/linux/man-pages/man8/ibqueryerrors.8.html

24. https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook
25. https://hpc.llnl.gov/documentation/tutorials/livermore-computing-commodity-clusters-overview-part-one
26. https://docs.oracle.com/cd/E26699_01/html/E26706/z40000fe1699290.html
27. https://docs.open-mpi.org/en/main/tuning-apps/networking/ib-and-roce.html
28. https://docs.open-mpi.org/en/main/release-notes/networks.html
29. https://openucx.readthedocs.io/en/master/running.html
30. https://networking-docs.nvidia.com/ibdiagnetutilityum/219/phy-diagnostics

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.