When Is a Server Ready? Evidence Sufficiency Under Partial Hardware Observability

A Cross-Vendor Empirical Study

Selim Kandaz

Draft manuscript v0.5.1 - 31 August 2026
Status: Working research draft. The empirical corpus comprises two deeply instrumented ASUS cases, a six-system HPE ProLiant DL360 Gen10 cross-vendor cohort with safety-gated escalation on two selected systems, and a retrospective corpus of 212 deduplicated runs across 23 identity-bound ASUS systems. Claims are limited to readiness decisions supported by the observed evidence; we do not estimate fleet failure rates, future reliability, or correctness of the proposed coordination mechanism.

Abstract

Operational readiness is not directly exposed by any single server interface. Operators infer it from partially overlapping observations such as management-controller state, firmware inventory, historical event logs, component telemetry, operating-system evidence, storage health, and bounded workload behavior. These observations differ in availability, retention, timing, and semantics and may disagree. We ask: **what evidence is sufficient to justify - or withhold - a scoped operational-readiness claim for a specific enterprise server?**

We model evidence sufficiency as **claim-dependent rather than absolute**. For an evidence set E and a scoped decision D, we define $S(E, D) = true$ when the available evidence supports the claim-specific predicates required for D and no unresolved blocking condition invalidates that claim. This formulation intentionally avoids a scalar health score and does not treat unobserved evidence as failure. It also exposes an asymmetry: a positive readiness claim may require multiple independent observations, whereas a sufficiently strong blocking observation may justify withholding the claim or stopping escalation without establishing that the hardware itself has failed.

We evaluate this model using two deeply instrumented ASUS server cases, a six-system HPE ProLiant DL360 Gen10 cohort, safety-gated escalation on two selected HPE systems, and a retrospective corpus of 212 deduplicated operational runs across 23 identity-bound ASUS systems. Across these studies, current and historical state diverged; observability changed with operating state; EMPTY, UNAVAILABLE, and QUERY_FAILED observations carried different evidentiary meaning; and software verdicts could diverge from independently re-observed physical outcomes. In HPE Phase 2, one system remained NOT_EVALUATED because the authorized path could not reach the active-validation layer. A second changed from Warning/Off to Critical/Starting after power-on, exposed two MapOutError DIMMs, and added 27 IML entries before workload execution; the experiment stopped by design.

Our central result is that readiness is a bounded claim justified by evidence, not a directly observable hardware property. Evidence sufficiency is therefore claim-dependent and asymmetric: additional evidence can increase confidence in the decision while decreasing assessed readiness.

Keywords: enterprise servers, operational readiness, hardware observability, evidence sufficiency, BMC, Redfish, IPMI, provenance, refurbishment, server validation

1. Introduction

Enterprise-server validation routinely ends in a small set of operational verdicts: PASS, FAIL, healthy, degraded, ready, or not ready. Such verdicts are necessary because operators and automation must eventually act. They are also lossy. **A verdict is a decision about a physical system; it is not the physical state itself.**

No single interface exposes that state completely. A Baseboard Management Controller (BMC) reports current management-plane observations. Firmware exposes configuration and inventory. Historical event logs retain selected past conditions. An operating system observes another subset of hardware state. Storage devices maintain their own health and lifetime information, while workload tests observe behavior only during the interval in which they execute. These observation planes differ in availability, retention, timing, and semantics, and their outputs can disagree.

This creates a problem of **partial hardware observability**. A missing event can mean that a log is genuinely empty, that a resource is unavailable, or that the query failed. A management controller can report a component as currently healthy while retaining evidence of earlier faults. A software workflow can report failure even when independent observation confirms that the intended physical firmware state was reached. A powered-off server can expose a different hardware view from the same server during POST. These distinctions determine not only what is observed, but also **what an operator is justified in claiming from the observation**.

We therefore ask a narrower question than whether a server is “truly healthy” or likely to remain reliable:

What evidence is sufficient to support a specific operational-readiness claim for a server, and what evidence is sufficient to withhold that claim or stop further validation?

Our key observation is that **evidence sufficiency is claim-dependent**. Let E denote the available evidence and D a scoped readiness decision. We write $S(E, D)$ for the proposition that E is sufficient for D. We define $S(E, D) = true$ when the available evidence supports the predicates required by D and no unresolved blocking condition invalidates the claim. We deliberately do not assign evidence a scalar health score: the same evidence set can be sufficient for one claim and insufficient for another.

This formulation also makes readiness decisions asymmetric. Establishing bounded workload readiness may require verified identity, supported configuration, relevant telemetry, and successful workload observations. By contrast, one sufficiently strong unresolved condition may justify withholding that claim before workload execution. Importantly, **withholding readiness is not equivalent to establishing hardware failure**. An unavailable validation path can leave a dimension NOT_EVALUATED; a contradictory observation can justify STOP_SAFETY; and an explicitly empty evidence source remains distinct from a failed query.

Our HPE experiments illustrate this distinction. In a read-only Phase 1 study of six ProLiant DL360 Gen10 systems, current management state and historical Integrated Management Log (IML) evidence were not equivalent. During a subsequent safety-gated escalation, one system could not reach the planned in-band validation layer through the authorized management path and therefore remained NOT_EVALUATED. A second system exposed a materially different state only after power-on: its aggregate state changed from Warning/Off to Critical/Starting, two DIMM records reported MapOutError, and the IML gained 27 entries before any workload executed. We stopped the experiment rather than forcing a workload verdict. **The additional evidence increased confidence in the decision while making the readiness assessment less favorable.**

The ASUS cases expose complementary forms of disagreement. A default SMART access path initially made available device-health information appear absent until a non-mutating SAT pass-through path was used. Clone-derived storage identity remained shared after higher-level appliance identities diverged. In another preserved case, a BIOS update physically reached its intended target while the controlling workflow reported failure because two software comparison paths interpreted the vendor-formatted version differently. Together, these cases show that observation path, identity binding, historical state, and interpretation can fail independently of the underlying component.

From these observations, we make four contributions. **First**, we introduce a claim-dependent model of evidence sufficiency for operational readiness that preserves categorical evidence states rather than collapsing them into a numerical health score. **Second**, we empirically characterize how current, historical, unavailable, state-dependent, and contradictory observations constrain readiness claims. **Third**, we evaluate these properties across two vendor ecosystems using deeply instrumented ASUS cases, a six-system HPE cohort with safety-gated escalation, and retrospective operational evidence. **Fourth**, we show that validation itself requires an escalation boundary: evidence collection may legitimately terminate with an unevaluated or withheld claim when the next observation layer is unavailable or unsafe.

The resulting contribution is intentionally narrower than a general theory of hardware reliability or decision-making. We do not estimate remaining useful life, derive fleet failure probabilities, or claim a universally sufficient test set. Instead, we study **which claims are justified by the evidence actually available at decision time**.

2. Research Questions and Claim Model

2.1 Research questions

The study is organized around four questions that are directly supported by the available experiments.

RQ1 - Evidence sufficiency. What evidence is sufficient to support a specific operational-readiness claim for enterprise hardware?

RQ2 - Incremental evidence. What additional information is contributed by historical, out-of-band, operating-state, and active in-band observations?

RQ3 - Conflict and absence. How should unavailable, contradictory, stale, empty, or state-dependent evidence constrain readiness claims and test escalation?

RQ4 - Transferability. Which evidence principles remain valid across different enterprise-server vendors and platforms despite differences in interfaces and telemetry semantics?

These questions are intentionally narrower than the earlier research program. They do not ask for fleet failure probabilities, an optimal economic test schedule, or a proof that one management protocol is superior to another. Those questions require larger controlled datasets.

2.2 Readiness as a bounded claim

We define operational readiness as a **scoped claim** about a particular physical system under an explicitly stated evidence window and intended use. Readiness is therefore not treated as a directly observable intrinsic variable.

Let $E=\{e_1,e_2,\ldots,e_n\}$ be the set of available evidence observations for a system, where each observation retains its source, time, semantic state, and physical identity binding. Let D denote a scoped decision or claim, such as:

identity verified for this physical server;

firmware state verified;

no relevant historical event was observed in the captured source;

bounded workload stability was observed during a defined interval;

ready for a specified operational handoff;

active escalation should stop because the next observation layer is unavailable or unsafe.

Evidence sufficiency is represented as:

$$S(E,D)=1 \Leftrightarrow \Phi_D(E)=true$$

where Φ_D is the logical predicate required for the specific decision D. The formalism is descriptive rather than probabilistic: it does not assign a percentage to readiness and it does not claim that all decisions use the same predicate.

For a positive readiness claim, the predicate is typically conjunctive:

$$\Phi_{ready}(E)=p_1(E) \wedge p_2(E) \wedge \ldots \wedge p_k(E)$$

where the required predicates can include identity continuity, configuration support, current telemetry, workload evidence, firmware state, and evidence completeness appropriate to the claim.

A stop/refusal decision can be structurally different:

$$\Phi_{stop}(E)=b_1(E) \vee b_2(E) \vee \ldots \vee b_m(E)$$

where a single sufficiently strong blocking condition can justify stopping escalation. This does **not** mean the hardware has necessarily failed. It means that continuing the planned validation is not justified under the observed state or authorized capability.

This asymmetry is the first major implication of the model: **the evidence required to justify readiness need not equal the evidence required to refuse escalation.**

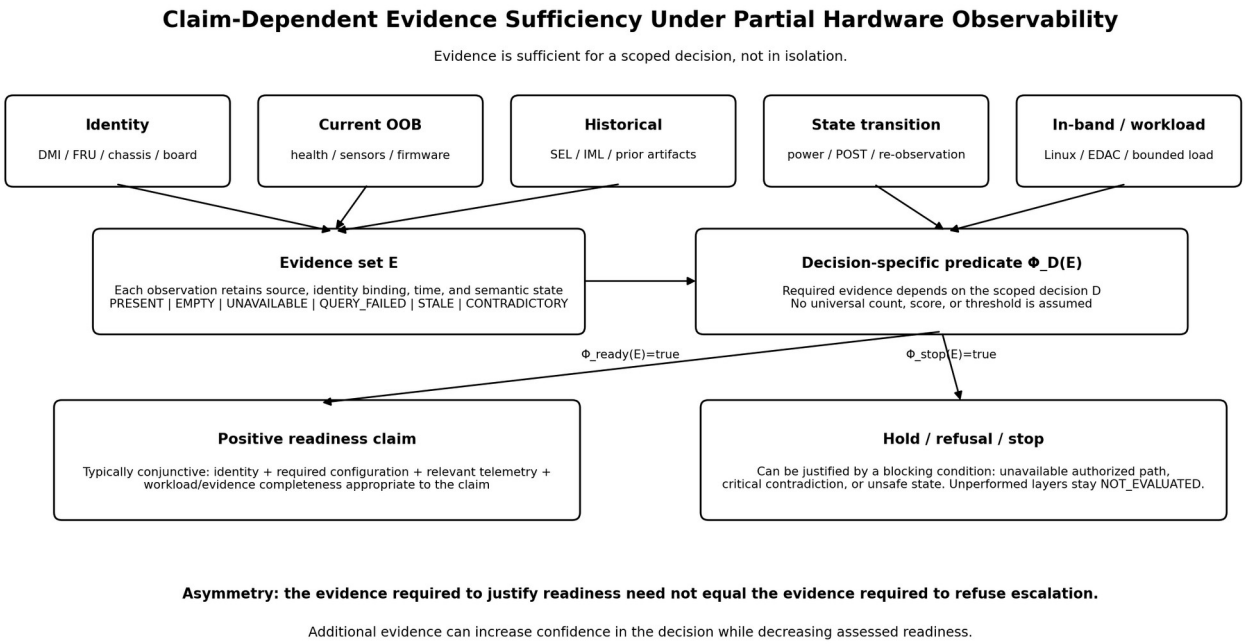

Figure 1. Claim-dependent evidence sufficiency under partial hardware observability.

2.3 Evidence semantics

Each evidence observation is preserved with semantics that distinguish the condition of the source from the apparent value returned by the source. The present study uses the following states:

PRESENT: the expected evidence was retrieved and parsed;

EMPTY: the source is valid and explicitly contains zero relevant events or members;

UNAVAILABLE: the required evidence is not exposed or cannot be accessed in the current authorized state;

QUERY_FAILED: the interface was addressed but the query itself failed;

INCOMPLETE: only part of the expected evidence was captured;

STALE: the value exists but does not reliably represent the current physical state;

AMBIGUOUS: evidence exists but its meaning cannot be resolved defensibly;

CONTRADICTORY: evidence sources support incompatible interpretations;

NOT_EVALUATED: the corresponding validation layer was not performed.

These states are not interchangeable. In particular, EMPTY is evidence; UNAVAILABLE is a limitation; and QUERY_FAILED is an observation-path failure. A zero value also requires state context: PowerConsumedWatts=0 while the host is powered off is not a measurement of active idle power, and a zero thermal field marked UnavailableOffline is not evidence of 0°C.

2.4 Readiness vector as a reporting descriptor

The earlier project used a categorical readiness vector:

$$R = \langle I, C, T, W, F, S, E, Co \rangle$$

for identity, configuration, telemetry, workload, firmware, security/provisioning, evidence completeness, and coordination. In this paper the vector is retained as a **reporting descriptor**, not as the sufficiency model itself. $S(E, D)$ determines whether the evidence supports a particular claim; the vector summarizes which dimensions are verified, partial, contradictory, unavailable, or not evaluated. Historical evidence remains a separate evidence layer rather than being compressed into one vector coordinate.

3. Related Work and Positioning

3.1 Health management, diagnosis, and telemetry

Electronics prognostics and health management (PHM) treats health as a changing condition inferred from tests, precursors, and life-cycle observations rather than as a one-bit property [1]. Multi-sensor information-fusion research similarly studies how heterogeneous observations can improve fault diagnosis and how missing or unreliable sources affect inference [2]. Server telemetry work demonstrates the value of continuous temperature, power, utilization, and error measurements for product-health insight [3], while SSD telemetry has been used at datacenter scale for endurance monitoring [7]. These traditions motivate multi-source observation but generally target diagnosis, prognosis, or continuous monitoring rather than a bounded acceptance decision for a specific server with uncertain history.

3.2 Field reliability and the limits of one-time tests

Large field studies of DRAM and flash show why present observations should not be conflated with future reliability. Schroeder et al. documented DRAM error behavior at scale [4]; Cheng et al. studied correlations between DRAM errors and server failure [5]; Meza et al. showed that flash reliability does not reduce to a simple wear measure [6]; and PinDrop showed that silent-data-corruption failures can emerge outside initial qualification windows [13]. The present study does not compete with those fleet analyses. Instead, it treats their results as a reason to scope readiness claims to the evidence window actually observed.

3.3 Management interfaces and provenance

IPMI and Redfish expose essential out-of-band state but differ in behavior and implementation [8]. The Redfish data model standardizes many resources while still allowing vendor-specific extensions [14]. The empirical results in this paper treat these interfaces as evidence channels rather than as ground truth. Provenance work such as ProvBench demonstrates the value of recording hardware, software, runtime, and result context for heterogeneous systems [9]. We extend that principle to physical acceptance decisions by retaining incorrect historical verdicts, later corrections, source semantics, and collection time as part of the evidence record.

3.4 Reuse and refurbishment

Server refresh and remanufacturing research shows that extending hardware lifetime can provide environmental benefits [10], [11]. Work on storage-media health metrics similarly argues that reuse confidence depends on interpretable health attributes rather than opaque lifetime labels [12]. Enterprise-server reuse adds an identity and evidence problem: a refurbisher may have only a short observation window, partial history, and multiple management layers that were not designed as a unified certification system.

3.5 Scoped research gap

In the consulted literature, we did not identify a single enterprise-server readiness method that jointly treats physical identity binding, out-of-band observations, historical state, state-dependent observability, in-band/workload evidence, explicit missing/empty/failed-query semantics, post-action re-observation, and immutable provenance as inputs to a claim-specific readiness decision. Neighboring work addresses important subsets of this problem. We therefore position the contribution as a cross-layer evidence method and empirical study, not as a claim that no related approach exists. A formal systematic literature review remains future work before publication-level novelty claims are frozen.

4. Study Design and Evidence Corpus

4.1 Three empirical pillars

The evaluation intentionally combines evidence with different strengths instead of pooling all observations into one statistical sample.

Pillar A - deep ASUS cases. Two ASUS enterprise servers were examined through management-plane and host-visible interfaces. The evidence includes DMI/SMBIOS, IPMI, Redfish, Linux EDAC/RAS, SEL state, sensor and power telemetry, SMART probing, bounded CPU workload, and preserved firmware/recovery artifacts. These cases are used for cross-layer causal observations rather than population statistics.

Pillar B - HPE cross-vendor cohort. Six HPE ProLiant DL360 Gen10 systems with iLO 5 were first examined read-only while powered off. Phase 1 collected identity, firmware, component population, current health, selected power/thermal telemetry, storage-controller state, security metadata, and IML history. Phase 2 selected HPE-02 and HPE-04 for a separately authorized ephemeral in-band escalation with explicit stop conditions.

Pillar C - retrospective operational evidence. The historical census reconstructed 212 deduplicated run identifiers, 211 non-empty runs, 93 runs with an explicit original disposition, 23 identity-bound physical ASUS systems, and 205 release-version tokens from more than 42,000 files. The corpus contains routine operational QA, controlled engineering validation, and unknown-context records. It is used for mechanism reconstruction and recurrence, not for failure-rate estimation.

4.2 ASUS deep cases

ASUS Case A - RS501A-E12-RS12U. Board family K14PA-U24-T Series, ASMB11, BIOS 2306 (Genoa), BMC firmware 1.02.38. The fresh study included a 90-second CPU calibration window together with identity, sensor, SEL, EDAC/RAS, power, and SMART observations.

ASUS Case B - RS720A-E13-RS12U. Board family K15PP-D24 Series, ASMB12, BIOS 1302 (Turin), BMC firmware 1.1.42. The same bounded CPU calibration was used, with additional preserved evidence from a prior firmware-lifecycle disagreement.

The 90-second workload is a calibration/observation window, not burn-in. It establishes only that no observed fault appeared in the captured channels during that bounded interval.

4.3 HPE Phase 1

All six HPE systems exposed Redfish 1.20.0 through iLO 5. System and Chassis model/serial identity agreed. Active memory capacity in the powered-off view ranged from 32 to 128 GiB. Current active DIMMs reported MultiBitECC and GoodInUse, while IML retained materially different historical records across systems. HPE-02 retained DIMM training and corrected-memory-threshold events even though current system/chassis health was OK. HPE-04 retained a historical critical No memory is available event. HPE-06 retained historical serial/product-ID and nonvolatile-memory-corruption records.

Several systems also exposed observation limitations. HPE-03/05/06 had stale device clocks in the 2001/2003 range, requiring independent collector time. HPE-02/05 returned zero/UnavailableOffline thermal fields in the powered-off state. All six reported PowerState=Off; therefore zero current power was an off-state observation rather than active idle consumption. Storage exposure also varied: a valid controller with zero drives was semantically different from a generic resource returning HTTP 400.

4.4 HPE Phase 2: planned escalation and safe termination

Phase 2 attempted to add an in-band Linux layer without writing internal storage. The protocol permitted transient power-on, once-only boot override, temporary read-only/RAM-backed media, bounded workload only after a verified HOST_READY predicate, re-observation, graceful shutdown, and media eject. Firmware, BIOS settings, iLO configuration, event logs, internal storage, users, and network configuration were not to be modified.

HPE-02 stopped at the access/precondition gate. Its authorized iLO capability did not provide the Virtual Media/Remote Console functionality needed by the protocol, and no safe PXE path was available. The machine remained powered off and workload evidence remained NOT_EVALUATED.

HPE-04 advanced through one temporary boot attempt. Management access remained available, but the host never reached a verified Linux/network predicate. Instead, the system changed from Warning/Off to Critical/Starting, remained in that state for approximately 17 minutes, exposed two MapOutError DIMM records, and increased IML from 37 to 64 entries. The experiment stopped before CPU or memory workload.

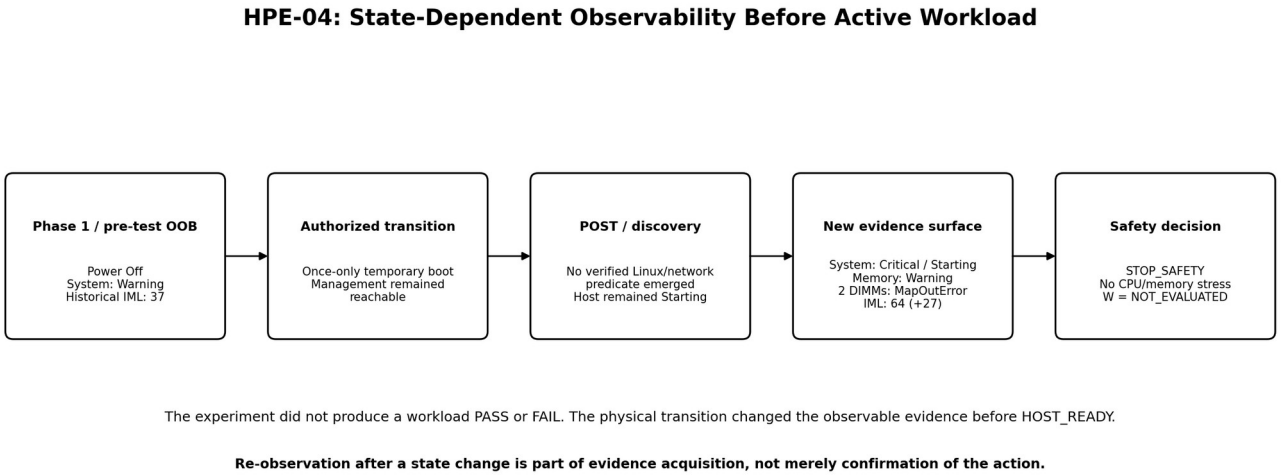

Figure 2. HPE-04 state-dependent observability and safety-gated termination.

4.5 Retrospective evidence and release archaeology

The retrospective corpus reconstructs failure-driven engineering changes across the automation lineage. Strong cases include cloned runner identity collisions; a missing platform contract before RS720 mutation; BMC recovery and readiness polling; evidence-delivery ordering; separation of empty SEL from endpoint failure; a vendor-native upload path that required persistent HTTPS connection reuse; and a physically successful BIOS transition that was initially classified as failure because version-comparison code mishandled a vendor suffix.

Original artifacts are preserved rather than rewritten after root-cause discovery. This matters because the historical incorrect verdict is itself evidence about the behavior of the validation system at that time.

4.6 Evidence provenance

Every important observation is associated with an evidence class such as PHYSICAL, ARTIFACT, CODE, VENDOR_DOC, PUBLIC_SOURCE, DESIGN, INFERENCE, or OPEN. Device time is also treated as evidence rather than assumed authority. When device clocks were stale or inconsistent, collector time was preserved independently. Publication-facing archives exclude credentials and reusable session material while retaining negative results and cryptographic checksums.

5. Findings

The Results section is organized by findings rather than by server. Each finding is supported by one or more concrete cases and is intentionally scoped to what those cases establish.

Finding 1 - No single representation was sufficient

Across the study, no individual representation captured every aspect required for a defensible readiness claim. The ASUS cases combined BMC sensors, host EDAC/RAS, SEL, power, SMART, identity, and workload because each source had a different observation boundary. HPE Phase 1 showed the same structure through a different vendor stack: iLO could establish identity, current component state, firmware, IML, and selected telemetry, but could not establish Linux bootability, kernel RAS state, or active workload behavior while the systems remained powered off.

This does not imply that every readiness decision always requires every available source. It implies that a source is sufficient only for claims within its observation scope. OOB evidence can be sufficient for identity or historical-log claims while remaining insufficient for an active workload claim.

Supported RQs: RQ1, RQ2, RQ4.

Finding 2 - Current state and historical state are non-equivalent

HPE-02 provided the clearest cross-vendor example. Its current system/chassis health was OK and active DIMMs reported GoodInUse, yet IML retained DIMM training events and a corrected-memory-threshold event. HPE-04 and HPE-06 contained different historical incidents that were not represented by a single present-tense health field.

The result is not that historical events prove current failure. The narrower conclusion is that current health and historical incident evidence are different evidence classes. A present-tense Health=OK observation cannot be interpreted as “no relevant historical hardware incident,” and a clean bounded workload would not erase historical evidence if one were later obtained.

Supported RQs: RQ1, RQ2, RQ3, RQ4.

Finding 3 - Observability can depend on physical operating state

HPE-04 showed that the observation surface itself changed after a physical state transition. In the powered-off view, the management interface exposed a limited active-memory picture. After power-on and POST discovery, the evidence package contained a much richer memory topology, aggregate memory Warning, two MapOutError DIMMs, Critical system state, and 27 additional IML entries. Linux never became available.

The finding is therefore stronger than the generic statement that “powered-off systems expose less telemetry.” It demonstrates that a previously collected OOB snapshot was not a stable substitute for post-transition re-observation on the same physical system. State transitions can reveal new evidence before the intended next validation layer is reached.

Supported RQs: RQ2, RQ3, RQ4.

Finding 4 - Missing, empty, negative, and failed observations are non-equivalent

Several cases show why a binary data-present/data-absent model is insufficient. A genuine zero-event SEL is EMPTY, not unavailable. A Smart Array controller with zero physical drives is a valid empty configuration. An HTTP 400 on a generic storage resource is QUERY_FAILED. HPE-02’s unavailable ephemeral boot path is UNAVAILABLE, not a hardware failure. Zero watts while a host is powered off is an off-state numerical value, not evidence of zero active idle consumption. Zero thermal data paired with UnavailableOffline is not a physical temperature.

The practical consequence is that evidence ingestion must preserve source state as well as returned value. Collapsing these cases into PASS, FAIL, or no-data removes information needed for later decisions.

Supported RQs: RQ1, RQ3, RQ4.

Finding 5 - Software verdicts can diverge from independently observed physical outcomes

The strongest case comes from the preserved ASUS RS720 BIOS lifecycle. The machine physically transitioned from BIOS 0601 to 1302, and post-reboot DMI confirmed the intended state. The controlling workflow nevertheless recorded failure because two comparison paths interpreted 1302 and 1302(Turin) as unequal. Later releases corrected the comparison logic; the original failed artifact remained immutable.

A second class of mismatch appeared in observation paths. Default SMART access behind a USB bridge suggested that SMART data were unavailable, while non-mutating SAT pass-through exposed usable underlying SMART information. In both cases the physical device state did not change; the software interpretation or observation path did.

These cases motivate independent re-observation: the action path or status field that performed an operation should not be the sole authority that certifies the resulting physical state.

Supported RQs: RQ1, RQ2, RQ3.

Finding 6 - Evidence acquisition can legitimately terminate before active testing

HPE Phase 2 was designed to add workload evidence but produced two valid terminal states before workload. HPE-02 ended with STOP_UNAVAILABLE: the authorized access path required for the active layer was absent, so W=NOT_EVALUATED. HPE-04 ended with STOP_SAFETY: the system became Critical/Starting with mapped-out memory before HOST_READY, so continuing to stress would have violated the experimental boundary.

Neither stop should be rewritten as a workload FAIL. The workload was not executed. Yet the stop decisions were themselves supported by evidence. This demonstrates the asymmetry in $S(E, D)$: evidence can be sufficient for a refusal/hold decision before it is sufficient for a positive readiness claim.

A related consequence is non-monotonic favorability. HPE-04 gained more high-confidence evidence after power-on, but its readiness descriptor became less favorable: configuration moved toward CONTRADICTORY while workload remained NOT_EVALUATED. **More evidence increased decision confidence while decreasing assessed readiness.**

Supported RQs: RQ1, RQ2, RQ3.

Finding 7 - Evidence principles transferred across vendors even when interfaces did not

The ASUS and HPE platforms differed substantially in BMC implementation, telemetry exposure, vendor extensions, storage behavior, historical logs, and authentication/management paths. The specific interfaces therefore did not generalize cleanly. The evidence principles did.

Across both vendors, identity had to be bound explicitly; current state and historical evidence had to remain separate; empty and unavailable observations had to remain distinct; action completion required re-observation; and unavailable evidence constrained the claim rather than automatically becoming failure. The HPE cohort also showed variation within one product family, reinforcing the point that model or BMC generation alone is not a sufficient observability contract.

The transferability claim is therefore about **evidence semantics**, not API equivalence or vendor performance.

Supported RQs: RQ4, with support from RQ1-RQ3.

5.8 Finding-to-evidence map

F1 - No single representation was sufficient. Primary support: ASUS deep cases and HPE Phase 1. Establishes that different layers expose different claim-relevant state; does not establish that every claim requires every layer.

F2 - Current and historical state are non-equivalent. Primary support: HPE-02/04/06 IML versus current health. Establishes that present health does not subsume retained history; does not show that historical events predict current failure.

F3 - State-dependent observability. Primary support: HPE-04 power transition. Establishes that the evidence surface can change with physical state; does not identify the exact root cause of the memory condition.

F4 - Missing, negative, empty, and failed observations are non-equivalent. Primary support: EMPTY SEL, HPE storage/thermal semantics, and HPE-02 access. Establishes that observation semantics affect valid inference; does not provide a universal vendor-independent error taxonomy.

F5 - Software verdicts can diverge from physical outcome. Primary support: RS720 BIOS and the USB SMART path. Establishes that interpretation/path can be wrong despite physical state; does not estimate the frequency of such errors in fleets.

F6 - A valid stop can occur before workload. Primary support: HPE-02 and HPE-04 Phase 2. Establishes that NOT_EVALUATED and stop decisions can be evidence-driven; does not establish HPE active-workload behavior.

F7 - Principles transfer while interfaces do not. Primary support: ASUS versus HPE. Establishes that core evidence semantics survived vendor change; does not imply universal behavior across all enterprise vendors.

6. Answers to the Research Questions

RQ1 - What evidence is sufficient to support a specific operational-readiness claim?

The study does not identify a universal minimum evidence count. Instead, it supports a claim-dependent answer: evidence is sufficient when the predicate required for the intended claim is satisfied and unresolved blocking conditions are absent. Identity claims can be supported by consistent identity sources without workload. A bounded workload claim requires an executed workload and corresponding observations. A general operational handoff can require multiple dimensions. The HPE Phase-2 stops show the inverse: evidence may be insufficient for readiness yet sufficient for a hold or stop decision.

Answer: sufficiency is claim-dependent, not absolute or scalar.

RQ2 - What additional information is contributed by different evidence layers?

Historical IML added incidents not present in current HPE health. Power transition added new memory and system-state evidence before Linux became available. In-band ASUS observation exposed EDAC/RAS and host identity not available from BMC alone. Active ASUS workload added bounded behavioral response and power observations. Retrospective artifacts added causality that current code or current hardware state could not reconstruct by themselves.

The HPE Phase-2 experiment did **not** complete the intended Linux/workload comparison, so the incremental value of HPE in-band EDAC/workload evidence remains open for the selected pair.

Answer: layers can add non-redundant information, but their incremental value is state- and access-dependent; the HPE in-band contribution remains incompletely measured.

RQ3 - How should unavailable, contradictory, stale, or state-dependent evidence constrain decisions?

They should remain explicit and constrain the scope of the claim. UNAVAILABLE prevents a claim that requires that layer; QUERY_FAILED indicates an observation-path failure rather than a healthy empty result; stale clocks require independent

collector time; contradictions require reconciliation or a narrowed claim; and a safety-critical state can stop escalation without converting unperformed tests into failure.

Answer: preserve uncertainty and disagreement rather than collapsing them into binary verdicts.

RQ4 - Which principles transfer across vendors and platforms?

The evidence semantics transferred more reliably than the interfaces. ASUS and HPE both required explicit identity binding, separation of current and historical state, distinction among empty/unavailable/failed observations, and scoped interpretation of active tests. Vendor-specific APIs, OEM fields, storage resources, telemetry availability, and management behavior differed substantially.

Answer: the cross-vendor contribution is a transferable evidence discipline, not a universal management API or universal component-health model.

7. Discussion

7.1 What does it mean for a server to be ready?

The empirical answer is deliberately bounded: **readiness is not a directly observable hardware property; it is a claim justified by evidence.** A server can be identity-verified while workload remains not evaluated. It can have current healthy DIMMs while retaining historical memory incidents. It can complete a bounded workload without that result becoming a lifetime reliability certificate. It can acquire stronger evidence and become less suitable for escalation.

This framing avoids two common errors. The first is overclaiming from a single green status. The second is treating any missing layer as failure. Both errors arise from compressing evidence before its semantics are understood.

7.2 Evidence sufficiency is asymmetric

The HPE Phase-2 outcomes reveal an asymmetry that is operationally important but easy to hide in PASS/FAIL systems. A positive readiness claim often requires a conjunction of conditions: identity continuity, supported configuration, relevant current telemetry, appropriate workload evidence, and sufficiently complete provenance. A stop decision can be triggered by a single blocking condition. That asymmetry does not mean a blocking condition proves component failure; it means the next escalation step is not justified.

This is the distinction between fail-safe practice and evidence semantics. A traditional workflow may simply mark the server FAIL when a severe condition appears. The method here preserves the stronger statement that is actually supported: STOP_SAFETY, together with the observed condition, while leaving unexecuted dimensions NOT_EVALUATED.

7.3 More evidence does not necessarily make a server more ready

Most validation workflows implicitly treat evidence accumulation as monotonic: every additional test is expected to move the system closer to a final PASS. HPE-04 shows why that assumption is wrong. The post-power evidence was more informative than the pre-power snapshot, but the system's readiness assessment became less favorable. Confidence in the decision increased because uncertainty decreased.

The objective of validation should therefore be **better justified decisions**, not favorable verdicts. This distinction is especially important in refurbishment, where commercial pressure can unintentionally reward workflows that collapse unavailable or contradictory evidence into a simple sale/no-sale label.

7.4 Implications for reuse and refurbishment

A reuse report built on this model would not state only "healthy" or "passed." It would state which claims are supported and which remain open. For example: identity verified; current telemetry supported; historical IML captured with prior incidents; bounded workload not evaluated; firmware verified; storage absent; evidence complete for the defined profile; future reliability not claimed.

Such a report is less concise than a health percentage but more auditable. It also creates a path toward later economic optimization: once enough standardized evidence is collected across a larger fleet, the marginal value and cost of each validation layer can be measured without first discarding its semantics.

7.5 Automation is a supporting implication, not the central contribution

The evidence study has consequences for automation because active observation can change physical state. A safe workflow should bind identity, observe preconditions, gate escalation, act only within authorization, wait for observable state predicates, re-observe the result, and preserve evidence. The compact engineering rule is:

Own -> Observe -> Gate -> Act -> Wait for State -> Re-observe -> Prove.

This paper does not present or validate a distributed coordination protocol. A time-bounded lease/fencing mechanism remains a candidate mitigation for the separately observed problem of concurrent mutation ownership, drawing on established lease concepts [15]. Its correctness, liveness, and failure behavior require a dedicated experiment and are outside the main contribution.

7.6 What the paper does not claim

The study does not establish fleet failure rates, remaining useful life, or a universal readiness threshold. It does not show that historical HPE memory events predict future failure. It does not rank HPE against ASUS. It does not prove that the eight-component readiness vector is the only valid representation. It does not prove that every enterprise platform exhibits the same state-dependent observability. It does not claim that a 90-second ASUS workload is burn-in. These limits are part of the result because $S(E, D)$ is meaningful only when D itself is scoped.

8. Threats to Validity

Small active sample. Only two ASUS systems received the full deep active profile. Their cases support mechanism-level observations, not population rates.

HPE cohort scope. The six HPE systems are all ProLiant DL360 Gen10 with iLO 5. Phase 1 was powered-off/OOB. Phase 2 selected only two systems and neither completed Linux workload. The HPE contribution therefore concerns evidence semantics, state transition, and safe termination rather than workload stability.

HPE configuration heterogeneity. Memory population, processor model, board details, controller presence, and firmware point versions differ. Same-family observability differences cannot be assigned to one causal factor without controlled isolation.

HPE-04 root cause unresolved. The MapOutError evidence and IML growth justify the safety stop but do not identify whether the cause was a DIMM, population rule, memory channel, board condition, or another factor.

Retrospective corpus bias. The 212-run historical corpus was created for operational QA and engineering, not prospectively for this research question. Procedures and software changed; repeated systems are common; metadata are incomplete; and negative-result retention may be uneven. The corpus is suitable for mechanism reconstruction, not unbiased incidence estimates.

No serial-bound historical Dell cohort. Dell-oriented software ancestry exists, but no serial-bound Dell CNServerOps run met the retrospective admission criteria. Dell cross-vendor conclusions therefore remain future work.

Bounded workload. The ASUS 90-second workload is a calibration/observation window. It cannot establish rare-defect coverage, thermal steady state, or future reliability.

Device-clock quality. Some HPE device clocks were stale. Collector time improves provenance but cannot reconstruct every historical event chronology.

Preliminary literature scoping. The related-work review identifies neighboring areas but is not yet a protocol-driven systematic review. Novelty claims remain intentionally conservative.

9. Future Work

The most important next experiment is not simply “more stress testing.” It is to test claim-dependent sufficiency prospectively on systems where the intended evidence layers can actually be reached. A future HPE in-band phase should use a separately authorized ephemeral boot path and a target whose pre-test state does not already require a safety stop. This would allow the incremental value of Linux EDAC/RAS and active workload to be measured relative to the OOB baseline.

A controlled Dell iDRAC cohort is also high value because Dell software ancestry exists but historical physical runs were not recoverable under the current census. Additional HPE generations, ASUS platforms, and component-level memory/storage studies can test whether the seven findings remain stable or require narrower wording.

The sufficiency model itself should also be tested prospectively. For each operational decision D, future work can define an explicit predicate Φ_D, run blinded or pre-registered validation profiles, and measure whether the predicate produces consistent decisions across technicians, vendors, and repeated observations. This is the path toward a stronger formal model; it should precede any attempt to assign numerical readiness scores.

Finally, a formal systematic literature review should freeze databases, search strings, screening criteria, and inclusion/exclusion decisions before strong novelty language is used in a submission.

10. Conclusion

A server does not directly expose readiness. It exposes evidence from imperfect representations. The central problem is therefore not how to find the one correct health field, but how to decide what can be claimed from the evidence that is actually available.

Across deep ASUS cases, six HPE systems, a safety-gated HPE escalation, and retrospective operational evidence, this study finds that no single representation is sufficient for all claims; current and historical state are distinct; observability can depend on physical operating state; missing, empty, failed, and negative observations are not equivalent; software verdicts can diverge from physical outcomes; evidence acquisition can terminate validly before workload; and the evidence principles transfer across vendors more reliably than the interfaces themselves.

The resulting answer to the paper's primary problem is partial but concrete. For an evidence set E and a scoped decision D, evidence is sufficient only when the predicate required for that decision is satisfied. There is no universal evidence count and no justified scalar health threshold in the present data. Positive readiness and refusal are asymmetric: the former can require multiple supported predicates, while the latter can be justified by a single blocking condition without converting unknown evidence into hardware failure.

Four sentences summarize the contribution:

Readiness is not a directly observable hardware property; it is a bounded claim justified by evidence.

Evidence sufficiency is claim-dependent rather than absolute.

The evidence required to justify readiness may exceed the evidence required to refuse escalation.

Additional evidence can increase confidence in the decision while decreasing assessed readiness.

For enterprise-hardware reuse and validation, the practical rule is therefore simple: collect enough independently bound evidence to support the claim being made, preserve what remains unknown, and stop rather than manufacture certainty when the next observation layer is unavailable or unsafe.

Data and Artifact Availability

The present manuscript is supported by four sanitized research packages with manifests and cryptographic checksums:

ASUS deep case-study / hardware-readiness archive: SHA-256 6225820112b5facbe63f8d81138c025ea8c4a50e8c6fa93f9b7cd480111804ed.

HPE Cross-Vendor Readiness Phase-1 archive: 37 files, SHA-256 5E1ABBA2C90EDF50346CD5ABF99966035E534A499B8FAADBE16AA3FEE3BC20BF.

HPE Cross-Vendor Readiness Phase-2 v2 archive: 273 files including its manifest, SHA-256 bb4558a44f868d84742221fefa6fe1200f8a1b3a5002e60bd79e6a7c6c79540b.

Historical Operational Evidence and Release Archaeology archive: 43 files, SHA-256 1194afe267fd6848c4c6e7b8920d6fa1ccdcb9fe0448e24845619cc4c1d59b5b.

The publication package is intended to include sanitized raw observations needed to reproduce the reported analyses, derived tables/datasets, analysis scripts, manifests, and checksums. Credentials, reusable authentication/session material, internal network identifiers, and customer/business-sensitive fields will be excluded or irreversibly redacted. Negative results, superseded claims, and contradictory historical artifacts will be retained when necessary to reproduce the scientific record.

References

[1] N. Vichare and M. Pecht, "Prognostics and health management of electronics," *IEEE Transactions on Components and Packaging Technologies*, vol. 29, pp. 222-229, 2008.

[2] T. Lin, Z. Ren, L. Zhu, Y. Zhu, K. Feng, W. Ding, K. Yan, and M. Beer, "A Systematic Review of Multisensor Information Fusion for Equipment Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1-48, 2025.

[3] F. Su, R. Kwasnick, J. Holm, W. Penner, H. Gartler, J. Boelter, Y. Zhou, B. Arbab, and M. Rothberg, "Product Health Insights Using Telemetry," *IEEE Design & Test*, vol. 41, pp. 56-64, 2024.

[4] B. Schroeder, E. Pinheiro, and W. Weber, "DRAM errors in the wild: a large-scale field study," in *Proceedings of ACM SIGMETRICS*, pp. 193-204, 2009.

[5] Z. Cheng, S. Han, P. Lee, X. Li, J. Liu, and Z. Li, “An In-Depth Correlative Study Between DRAM Errors and Server Failures in Production Data Centers,” in *Proceedings of the 41st International Symposium on Reliable Distributed Systems (SRDS)*, pp. 262-272, 2022.

[6] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, “A Large-Scale Study of Flash Memory Failures in the Field,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, pp. 177-190, 2015.

[7] O. O. Medaiyese, F. Lin, H. Dixit, R. Mishra, A. Baglioni, L. Silva, M. Elkin, A. Ilyashenko, G. Safaryan, D. S. Ajravat, X. Jiao, and V. Parekh, “Hardware Telemetry at Scale: A Case Study on SSDs Endurance Monitoring in Datacenters,” in *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*, pp. 147-152, 2025.

[8] E. Hojati, Y. Chen, and A. Sill, “Benchmarking Automated Hardware Management Technologies for Modern Data Centers and Cloud Environments,” in *Proceedings of the 10th IEEE/ACM International Conference on Utility and Cloud Computing*, 2017.

[9] F. Liu, M. Belgin, N. Zhang, K. Manalo, R. Lara, C. P. Stone, and P. D. Manno, “ProvBench: A performance provenance capturing framework for heterogeneous research computing environments,” *Concurrency and Computation: Practice and Experience*, vol. 34, 2022.

[10] R. Bashroush, N. Rteil, R. Kenny, and A. Wynne, “Optimizing Server Refresh Cycles: The Case for Circular Economy With an Aging Moore’s Law,” *IEEE Transactions on Sustainable Computing*, vol. 7, pp. 189-200, 2022.

[11] F. Ardente, L. Talens Peiro, F. Mathieux, and D. Polverini, “Accounting for the environmental benefits of remanufactured products: Method and application,” *Journal of Cleaner Production*, vol. 198, pp. 1545-1558, 2018.

[12] R. Kenny, J. Hands, and N. Hayhurst, “From Waste to Resource: How Standardized Health Metrics Can Accelerate the Circular Economy in Storage Media,” *European Journal of Electrical Engineering and Computer Science*, 2024.

[13] P. W. Deutsch, H. Dixit, G. Vunnam, C. Moran, E. Ozer, and S. Sankar, “PinDrop: Breaking the Silence on SDCs in a Large-Scale Fleet,” in *2026 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 1-14, 2026.

[14] DMTF, *Redfish Data Model Specification*, DSP0268, 2025.3.

[15] C. G. Gray and D. Cheriton, “Leases: an efficient fault-tolerant mechanism for distributed file cache consistency,” *ACM SIGOPS Operating Systems Review*, vol. 23, pp. 202-210, 1989.