EAD RESEARCH PROGRAMME · EAD-2026-05 · VERSION 2.0

The Effective Life Problem

Hardware Obsolescence, the Byproduct Economy, and the Resource Allocation Question in Agentic AI Infrastructure

Author Shubham Agarwal, Founder and Chief Executive Officer, triNetra Research

Series EAD-2026-05

Status Working Paper

Date September 2026

Version 2.0

Conflict of Interest Statement: The author declares no financial or non-financial conflicts of interest. No external funding was received. All empirical material derives from public regulatory filings, vendor technical disclosures, preprint research, and independent industry reporting, cited throughout and characterized by source type in Section 8 and the References.

Keywords: external AI dependence, hardware obsolescence, effective useful life, embodied resource cost, secondary markets, byproduct economy, resource attribution, multi-agent systems, resource-adjusted ROI

ABSTRACT

Cost accounting for AI infrastructure typically treats a hardware unit as a single capital input, acquired once and consumed by the workload it was purchased for. This paper argues that treatment omits two related facts. First, a GPU or server can remain physically functional for years after it stops being the economically competitive choice for the specific AI workload it was acquired for, because successive hardware generations improve performance per watt and cost per token faster than physical components fail. Second, hardware displaced from that original workload does not typically stop generating economic or material value; it is commonly resold, redeployed to less demanding computing roles, or eventually recovered for components and materials. Treating a single "useful life" figure as sufficient, and treating an asset's first application as its only economically relevant use, understates and overstates AI's resource burden respectively, and the two errors partially offset in ways current accounting frameworks do not make visible.

This paper makes three contributions. First, it develops a temporal model of AI hardware useful life that distinguishes physical durability from workload-specific economic competitiveness, and introduces Effective Hardware Life Cost (EHLC) as a conceptual counterpart to Hardware Capital Cost. Second, it proposes a two-ledger accounting structure that keeps a primary-product ledger, tracking the resource-to-return chain for an AI system's original deployment, separate from a byproduct ledger tracking value the same physical assets generate after leaving that deployment. Third, it applies this structure to the resource-allocation question raised by increasingly compute-intensive multi-agent architectures, reframing the relevant comparison from token cost per task to the marginal value of directing constrained hardware toward additional primary-product compute versus toward the secondary uses the byproduct ledger tracks.

The paper is grounded in public evidence: divergent 2025 depreciation-schedule revisions at Amazon and Meta, vendor-disclosed generational performance and embodied-carbon data, and reported secondary-market pricing for retired accelerators. It does not estimate an empirical allocation of embodied resource cost between primary and secondary use; it identifies the allocation problem, proposes candidate allocation bases, and leaves their empirical resolution to subsequent research. The paper's scope is conceptual and structural rather than econometric, and its limitations are stated explicitly in Section 8.

TABLE OF CONTENTS

1. Introduction
2. Positioning Within the EAD Programme
3. A Temporal Model of Hardware Useful Life
4. Evidence: Depreciation Divergence and the Workload-Specific Obsolescence Hypothesis
5. The Byproduct Economy and the Attribution Problem
6. A Two-Ledger Accounting Structure
7. Application: The Multi-Agent Resource Allocation Question
8. Limitations
9. Discussion and Implications
10. Open Questions for Subsequent Research
11. Conclusion
12. References

1. Introduction

A physical accelerator purchased for frontier AI work can remain fully operational for years after it stops being the correct hardware for that work. Independent industry reporting places the compression of AI-specific refresh cycles inside frontier training and serving fleets at roughly eighteen to thirty-six months, well short of the five-to-seven-year horizon those same accelerators were historically expected to serve [1]. Nothing in that shift implies the hardware has failed. It implies that a newer generation now delivers more useful computation per watt, per dollar, and per unit of floor space, and that continuing to run the older generation for the same workload has become the more expensive choice.

That single observation creates two distinct accounting problems, and this paper's central argument is that they must be addressed together rather than separately, because each one corrects an error the other would otherwise leave standing.

The first problem is a cost-understatement problem. A framework that records only how much hardware capacity a workload required, its Hardware Capital Cost, does not record how long that capacity remained the economically correct capacity for the workload it was bought to serve. A unit that becomes workload-specific-obsolete after eighteen months has, in an important sense, delivered less than its nominal purchase price implies, even though it has not failed and continues to have residual value elsewhere. Omitting this understates the true resource cost of operating at the frontier.

The second problem is a cost-overstatement problem, and it points in the opposite direction. When a unit is displaced from frontier AI work, it does not generally exit the economy. Reported secondary-market activity for retired data-center accelerators indicates displaced hardware is

commonly resold, redeployed to less demanding inference workloads, or used in research and conventional computing roles before eventual component or material recovery [2][3]. A cost or environmental-footprint accounting that assigns the entirety of a unit's embodied material and manufacturing burden to its first, frontier-AI deployment, without crediting these subsequent uses, overstates how much of that burden should be attributed to the initial AI application.

Addressing only the first problem would make AI infrastructure investment look more expensive than a full-lifecycle view supports, by treating displaced hardware as if it disappeared rather than continuing to generate value elsewhere. Addressing only the second would understate the genuine cost of chasing frontier performance, by treating a unit's useful contribution as extending across its full multi-year physical life even where its competitive usefulness for the acquiring workload ended much earlier. This paper argues that a lifecycle accounting of AI hardware needs both corrections simultaneously.

This paper makes three contributions. First, it develops a temporal model of AI hardware useful life (Section 3) that distinguishes an accelerator's physical durability from its workload-specific economic competitiveness, and proposes Effective Hardware Life Cost (EHLC) as a conceptual complement to the conventional Hardware Capital Cost figure. Second, it proposes a two-ledger accounting structure (Section 6) that keeps the resource-to-return chain for an AI system's original deployment separate from a ledger tracking value the same physical assets generate after leaving that deployment, so that neither the primary cost nor the secondary value is obscured by netting one against the other. Third, it applies this structure to a live design question, the resource cost of increasingly compute-intensive multi-agent architectures (Section 7), reframing it as a question about the marginal value of directing constrained hardware toward additional primary compute versus toward the secondary uses the byproduct ledger makes visible.

Scope. This paper is conceptual and structural. It does not estimate, for any real fleet or firm, what share of an asset's embodied resource footprint should be allocated to its primary AI use; Section 6 identifies five candidate allocation bases without selecting among them, and Section 10 states why that selection is left to subsequent work. It does not present a life-cycle assessment, a causal identification of workload-specific obsolescence, or an investment recommendation. Section 8 states these and other limitations explicitly, organized by the kind of limitation involved rather than treated as a single caveat.

The paper proceeds as follows. Section 2 states briefly how this paper relates to and extends prior work in the same research programme, without requiring that prior work to be read first. Section 3 develops the temporal model and EHLC. Section 4 assembles the evidence for treating obsolescence as workload-specific rather than physical. Section 5 documents the byproduct economy and states the attribution problem it creates. Section 6 proposes the two-ledger structure. Section 7 applies the framework to multi-agent architecture. Section 8 states limitations. Sections 9 through 11 discuss implications, state open questions, and conclude.

2. Positioning Within the EAD Programme

This paper is the fifth in triNetra's External AI Dependence (EAD) research programme. A reader unfamiliar with the earlier papers does not need them to follow the argument here; this section states only what is inherited.

The programme's first paper proposed a material-to-return chain for AI infrastructure investment, in short: material inputs become hardware, hardware becomes compute, compute becomes inference and agentic output, and output is claimed as return on investment [4]. That chain recorded how much hardware a workload required, but not how long the acquired capacity remained the economically correct capacity for that workload, nor what happened to it afterward. This paper's contribution is to insert both a temporal dimension and a lifecycle-branching structure into that chain, which revises its central logic rather than adding a peripheral observation to it.

Three further concepts from the programme recur in this paper and are stated here so the paper is self-contained. The Judgment Layer, introduced as the institutional architecture an organization uses to synthesize operational signals into accountable decisions, is relevant because the choice between directing a displaced asset toward secondary value or leaving that decision to default is itself a judgment-layer decision [5]. The Infrastructure Loop and Economic Participation Framework, proposed as collective-level responses to infrastructure dependency, are not directly extended here but share this paper's premise that infrastructure is a resource whose allocation, not merely its acquisition, is the relevant unit of analysis [6]. PaaF, the programme's structural-interpretation method, is used explicitly in Section 3.5 to state this paper's Domain, Scale, Use Case, and Perception.

3. A Temporal Model of Hardware Useful Life

3.1 Extending the material-to-return chain

The prior chain read, in abbreviated form: material into hardware into compute into inference into agentic output into economic outcome into return on investment. This paper inserts a temporal stage between hardware and compute: material into hardware into effective life into compute into inference into agentic output into economic outcome into return on investment.

"Effective life" is deliberately not a single figure. Collapsing several distinct lifetimes into one useful-life number is, this paper argues, the specific modeling choice that produced the depreciation-schedule divergence documented in Section 4. Five lifetimes are distinguished:

- **Physical lifetime**: how long the unit continues to function, independent of whether any workload wants to run on it.
- **Economic lifetime**: how long the unit's operating cost remains below the value it generates in any workload, not necessarily its original one.

- **AI-specific useful lifetime**: how long the unit remains competitive for the specific frontier or near-frontier AI workload it was acquired for. This is the quantity industry reporting places at roughly eighteen to thirty-six months in frontier fleets [1], while physical lifetime is unchanged.
- **Residual or secondary lifetime**: how long the unit continues to generate value once displaced from its original workload into a less demanding one.
- **End-of-life and recovery**: the point at which the unit is decommissioned for component or material recovery rather than continued operation.

3.2 Effective Hardware Life Cost: a conceptual formulation

Hardware Capital Cost records what capacity was acquired. It does not record how long that capacity remained the economically correct capacity for the workload it was acquired for. This paper introduces Effective Hardware Life Cost (EHLC) to make that second quantity explicit.

EHLC is defined here as a conceptual formulation, not an empirically estimated quantity. The paper does not have, and does not claim to have, the fleet-level utilization, revenue, and disposition data such an estimate would require; Section 8 states this as a data limitation. The purpose of stating EHLC formally is to make clear what would need to be measured, not to imply it has been.

Conceptually, EHLC for a given hardware unit can be expressed as a function of:

```
EHLC = f( Acquisition Cost, Expected Useful Life, Realized Utilization,
Performance Relative to Current-Generation Hardware,
Operating Energy and Cooling Cost, Timing of Replacement, Residual Value at
Displacement )
```

Read in plain terms: EHLC is intended to capture the gap between what a unit cost to acquire and operate and what it actually delivered before its AI-specific competitiveness ended, adjusted for whatever residual value it retained at the point of displacement. Hardware Capital Cost asks how much capacity was required. EHLC asks how long that capacity remained the economically correct capacity for the workload, and at what operating cost, before a newer generation made continued use the more expensive choice.

A worked numerical estimate of EHLC for any real deployment would require fleet-level utilization data, realized (not list) acquisition pricing, energy costs specific to the site, and disposition outcomes, none of which are public for any single operator at the granularity this formulation would require. This paper's contribution at this stage is the conceptual distinction and the variables it depends on, not a populated estimate.

3.3 The branching lifecycle after AI-specific useful life

The prior chain implicitly assumed a single life: material becomes hardware, hardware serves one workload, and the framework's accounting for that material ends when the workload ends. This paper replaces that assumption with a branching structure. After AI-specific useful lifetime ends, an asset can be resold, redeployed within the same organization to a less demanding workload, or shelved; each of those branches can be followed by further branches, secondary AI inference, conventional enterprise computing, component recovery, material recycling, before the asset's material content finally exits the economy.

The practical consequence is that a framework asking "how much material did this AI workload consume" is asking an incomplete question. The more complete question is "what fraction of this asset's total resource footprint, across every use it will have, should be attributed to this particular workload." That is an allocation problem, not a subtraction problem, and Section 6 proposes a structure for keeping that allocation visible rather than resolving it by assumption.

3.4 Primary path and branching lifecycle, stated as a figure

The primary path this paper's material-to-return chain describes is linear. What happens after AI-specific useful life ends is not, and the branching is the point; the figure below should not be read as implying every asset follows the same subsequent sequence.

```
PRIMARY PATH (single sequence, applies while the asset serves its acquiring workload)

Material -> Infrastructure -> Hardware -> Effective Hardware Life ->
Compute -> Inference -> Agentic Operation -> Human Input ->
Productive Output -> Economic Value -> Resource-Adjusted ROI

AFTER AI-SPECIFIC USEFUL LIFE ENDS (branching, not a fixed sequence)

Hardware -> [ redeployment | resale | secondary AI inference |
              conventional computing | component recovery |
              material recovery ]   (any of these, not all, and not in order)
```

3.5 Structural interpretation

Restating this paper's contribution in triNetra's PaaF terms: Domain is the physical compute substrate underlying AI infrastructure. Scale runs from an individual asset's lifecycle through fleet-wide and market-wide secondary-economy effects. Use case is cost-stack accounting for AI infrastructure investment and the strategic choice between directing resources toward primary or secondary deployment. Perception is the difference between viewing a retired GPU as a sunk cost, the view an accounting framework stopping at first use implies, and viewing it as a still-active asset that has changed markets, the view this paper's branching structure makes visible.

4. Evidence: Depreciation Divergence and the Workload-Specific Obsolescence Hypothesis

This section states the evidence for treating AI hardware obsolescence as workload-specific rather than physical, and is explicit about what that evidence does and does not establish.

4.1 Generational performance gaps

The scale of generational improvement in metrics relevant to frontier AI workloads is, on its own, sufficient to motivate compressed refresh cycles without any assumption about component failure. Vendor and industry-analyst comparisons report that NVIDIA's H100 delivered roughly three times the performance per watt of the preceding A100 generation, and that the Blackwell-generation B200 and GB200 deliver a further generational gain of roughly two to two-and-a-half times over the H100 on a like-for-like precision basis, with the GB200 platform reporting a 47 percent improvement in TFLOPS per GPU-watt over the H100 [7]. These are vendor-reported and industry-analyst figures rather than independently reproduced benchmark results, and the reader should weight them accordingly.

A further figure requires more caution. Vendor-published inference-throughput comparisons for the GB200 platform report up to a 30-times inference-speed advantage over the H100 under specific benchmark conditions [7]. That comparison pairs a newer numeric precision (FP4) against an older one (FP8) and a single fast interconnect domain against hardware distributed over a slower one; it is a best-case platform comparison under vendor-selected conditions, not a like-for-like accelerator comparison, and this paper does not rely on it. Even the more conservative like-for-like figures imply that a fleet purchased two generations ago can require two to four times the power, floor space, and operating cost per unit of useful inference work as the current generation, without any of the older units having failed or physically degraded.

4.2 Depreciation-schedule behavior at major cloud operators

Hyperscaler depreciation schedules are accounting judgments about expected economic service life, not physical durability measurements, and their recent history is informative precisely because it has not moved uniformly. Amazon, Microsoft, and Google extended server useful-life assumptions from three to four years in 2020 and, by 2023 to 2024, to six years across the major cloud operators; independent analysis attributes a combined reduction in the three operators' annual depreciation expense from an estimated 39 billion dollars to 21 billion dollars in 2024 to this extension [8][9]. That extension reflects the reality that a large share of hyperscaler fleet capacity runs conventional cloud and storage workloads, where six or more years of physical service life is genuinely available.

Amazon's own third-quarter 2025 filing with the U.S. Securities and Exchange Commission states that, effective January 1, 2025, Amazon changed its estimate of the useful lives of a subset of its servers and networking equipment from six years to five years, citing "the increased pace of

technology development, particularly in the area of artificial intelligence and machine learning" [10]. For the nine months ended September 30, 2025, the filing states this change increased depreciation and amortization expense by 889 million dollars and reduced net income by 677 million dollars [10]. Meta, over a comparable period, extended the useful life of most of its servers and network assets to five and a half years [9].

This divergence is evidence that the two firms formed different expectations about the economic service life of comparable asset categories over the same window. It is evidence of heterogeneity, not proof of a specific mechanism. Section 8.1 states the alternative explanation, that the divergence reflects financial-reporting incentives rather than a genuine signal about workload-specific obsolescence, and why this paper treats that alternative as only partially addressed rather than ruled out.

4.3 Secondary-market pricing

Reported secondary-market pricing for retired data-center accelerators is broadly consistent with hardware retaining substantial value once judged uncompetitive for frontier AI use. Multiple industry sources report used NVIDIA A100 accelerators, two generations behind current frontier hardware, retaining roughly 60 to 80 percent of original purchase price in resale markets [3][11]. A separate, single-source report of GPUs from expired cloud contracts being rebooked at up to 95 percent of original pricing when redeployed into less demanding roles could not be independently corroborated in this research pass and is flagged in Section 8.2 rather than presented as established.

Hardware retaining a substantial majority of its purchase-price value after being judged uncompetitive for its original purpose is consistent with an obsolescence condition specific to that purpose rather than a statement about the physical unit. This is consistent with, not proof of, the workload-specific obsolescence hypothesis; Section 8.1 states why depreciation-schedule and resale evidence alone cannot fully distinguish that hypothesis from the financial-engineering alternative.

5. The Byproduct Economy and the Attribution Problem

5.1 Two separate propositions

This paper depends on distinguishing two propositions that are easy to conflate. The first is empirical: displaced AI hardware commonly retains resale or redeployment value. The second is methodological: therefore, an accounting framework that assigns an asset's full embodied resource footprint to its first application may be incomplete, and a defensible allocation mechanism across an asset's subsequent uses is required. The first proposition is supported by the evidence in Section 4.3. The second does not follow automatically from the first; it is this paper's own methodological argument, stated and defended in this section, not a finding the evidence alone establishes.

5.2 Where displaced hardware goes

Reporting on the secondary market for AI accelerators describes several distinguishable stages rather than a single handoff: resale to external buyers seeking pre-owned data-center GPUs, redeployment within the original operator's own fleet toward less latency-sensitive or smaller-model inference workloads, use in research, simulation, or rendering roles that do not require frontier throughput, and eventual disposition through formal IT asset disposition and recycling channels once no further compute role remains economical [2][3]. Industry reporting also attributes part of secondary-market demand growth to 2025 tariff policy raising new-component costs, though the magnitude reported (a 20 to 40 percent increase) rests on a single industry source this paper could not independently corroborate and is accordingly flagged in Section 8.2 [2].

5.3 The material case for separate tracking

The case for tracking secondary use separately from primary use is strongest at the embodied-material level. NVIDIA's own 2025 Product Carbon Footprint disclosure for the HGX H100 system reports a total cradle-to-gate impact of 1,312 kilograms of CO2-equivalent per unit, with high-bandwidth memory contributing 42 percent of that impact, integrated circuits 25 percent, and thermal management components 18 percent [12]. This figure is vendor-reported (NVIDIA's own disclosure, independently cited as a validation benchmark in at least one preprint life-cycle study of a comparable accelerator [12]), not an independently audited third-party measurement, and is characterized as such throughout this paper.

Semiconductor fabrication is separately water-intensive. A large fabrication facility can use up to 38 million liters of ultrapure water per day, a figure comparable to a mid-sized city's daily consumption [13]. TSMC's own 2022 Sustainability Report states the company used approximately 35 billion gallons, roughly 132 billion liters, of ultrapure water across its global operations in 2022, a 21 percent increase over 2021 [14]; this figure replaces an earlier, less precisely sourced withdrawal-and-discharge estimate and is now cited directly to TSMC's own company disclosure.

A frequently cited industry estimate holds that global integrated circuit production accounts for approximately 185 million tons of CO2-equivalent emissions annually [15]. This figure recurs across several industry publications (imec, Semiconductor Digest, and others), but this paper's research pass could not trace it to a single primary methodology document with an auditable calculation. It is reported here as a widely circulated industry estimate, not a verified primary figure, and the author should confirm its original source before this paper relies on it in any high-stakes citation.

Charging the entirety of an asset's embodied footprint against a single eighteen-to-thirty-six-month frontier deployment, when the same physical memory, integrated circuits, and thermal hardware may go on to serve further economic roles before final recovery, is the specific overstatement the two-ledger structure in Section 6 is designed to make visible and correctable. Correcting this

overstatement does not reduce the absolute resource footprint of manufacturing the hardware in the first place; it changes which economic activity should be charged for which share of that already-incurred footprint. This distinction is restated in Section 6.4 because it is easy to lose.

6. A Two-Ledger Accounting Structure

6.1 Purpose

This paper proposes keeping primary-product accounting and byproduct accounting in two separate ledgers rather than netting one against the other inside a single row. The purpose of this separation is not to reduce AI's apparent resource footprint. It is to avoid assigning the entire lifecycle burden of a physical asset to one economic use when the same asset subsequently generates value, and consumes further resources, in other economic contexts. Keeping the ledgers separate also lets the byproduct ledger be examined as a research object in its own right, since the allocation question it raises, addressed in Section 6.3, is unresolved and independently interesting.

6.2 Table 1: primary product, material to return

Category	Aggregated metric	Derivation	Ultimate output
Material	Material Resource Input	Physical resources embodied in required infrastructure	Material footprint
Infrastructure	Infrastructure Requirement	Datacenter, power, cooling, network, and facility capacity	Infrastructure footprint
Hardware	Productive Hardware Capacity	Accelerators, memory, servers, and networking required	Hardware footprint
Hardware lifecycle	Effective Hardware Life Cost (EHLC)	Acquisition cost, expected useful life, realized utilization, and performance relative to current-generation hardware, through to displacement and residual value (Sec. 3.2)	Lifecycle-adjusted capacity
Energy	Operating Energy	Electricity and cooling attributable to the workload	Energy footprint
Compute	Compute Requirement	Actual compute capacity consumed	Compute footprint
Inference	Inference Requirement	Model calls, tokens, and reasoning operations	Inference workload
Agentic operation	Agentic Resource Requirement	Parallel agents, orchestration, context, retrieval, verification, and retries	Agentic workload
Human	Human Resource Requirement	Development, supervision, exception handling, and maintenance	Human input
Outcome	Successful Productive Output	Useful work actually delivered	Output
Value	Economic Value Generated	Revenue, cost reduction, productivity gains	Economic value
Return	Resource-Adjusted ROI	Economic value divided by total attributable resource cost	Primary economic result

Each row's aggregated metric names the quantity Table 1 tracks; the derivation column states what that quantity is built from; the ultimate output column states what the row contributes to the chain in Section 3.1. Resource-Adjusted ROI differs from a conventional ROI calculation specifically in that its denominator is intended to include EHLC rather than only Hardware Capital Cost, which is the mechanism by which this table's central correction (Section 3.2) enters the return calculation.

6.3 Table 2: the byproduct and secondary economy

Table 2's ten categories are grouped below into four kinds of residual, because treating all ten as a single undifferentiated "byproduct" concept would understate real differences in how each is measured, transferred, and governed.

A. Physical-asset residuals (the retired unit itself and its constituent components)

Category	Aggregated metric	Source	Potential secondary use	Commercialization mechanism	Value metric
Retired hardware	Secondary Hardware Value	Displacement from frontier AI workloads	Secondary compute, resale, leasing	Resale or redeployment	Economic value
Components	Component Recovery Value	Shelved or decommissioned equipment	Parts and refurbishment	Secondary market	Recovery value
Materials	Recovered Material Value	End-of-life equipment	Recycling and material recovery	Commodity or material market	Material value

B. Infrastructure and capacity residuals (facility and network capacity freed by workload migration, not the physical asset itself)

Category	Aggregated metric	Source	Potential secondary use	Commercialization mechanism	Value metric
Repurposed compute	Secondary Compute Capacity	Hardware moved off frontier workloads	Smaller-model or conventional inference	Capacity reallocation	Compute-hours and value
Infrastructure	Residual Infrastructure Value	Datacenter or network capacity freed by workload migration	Other workloads	Reallocation or leasing	Capacity value
Energy and capacity	Recoverable Capacity Value	Capacity released through workload optimization	Alternative computational use	Reallocation	Capacity value

C. Intangible residuals (non-physical outputs of primary operation that retain value independent of the hardware)

Category	Aggregated metric	Source	Potential secondary use	Commercialization mechanism	Value metric
Data	Secondary Data Value	Data generated or processed during primary operations	Analytics, training, knowledge products	Licensing or reuse	Economic value
Models	Residual Model Value	Older or smaller models displaced by newer releases	Edge, enterprise, or specialized inference	Licensing or deployment	Economic value
Knowledge and artifacts	Reusable Knowledge Value	Agent-generated intermediate outputs	Future workflows or products	Reuse	Economic value

D. Human-capability residuals (labor capacity freed by the primary deployment, included because the material-to-return chain in Table 1 treats Human Resource Requirement as a primary-side input, and its release is the direct counterpart on the byproduct side)

Category	Aggregated metric	Source	Potential secondary use	Commercialization mechanism	Value metric
Human capability	Redeployed Human Value	Labor freed by process change or automation	Higher-value work	Redeployment	Productivity and value

This grouping is a conceptual clarification, not a claim that the four categories require different allocation mechanisms; Section 6.4 states that the allocation question is open for all four alike, and a future research cycle should determine whether the allocation basis in fact needs to differ by category.

6.4 The allocation question

Table 2's rightmost columns are deliberately not pre-filled with assumed figures. What fraction of an asset's total lifecycle value should be allocated back against its primary AI use is a methodological choice this paper identifies but does not resolve. Five candidate allocation bases are identified: compute utilization, useful compute delivered, revenue generated, operating hours, and physical resource depletion. Each would assign a different share of the embodied footprint documented in Section 5.3 to the primary AI workload. As stated in Section 5.3, resolving this allocation question would not change the absolute resource footprint already incurred in manufacturing the hardware; it would change how that already-incurred footprint is distributed across the economic activities that subsequently draw on the same physical asset. Choosing among the candidate bases is left to subsequent work in this programme, stated further in Section 10.

7. Application: The Multi-Agent Resource Allocation Question

7.1 The resource cost of additional agentic structure

With both ledgers in view, the industry's direction toward more agents, more inference calls, and more orchestration overhead per task can be restated as a resource-allocation choice rather than only a capability choice. A controlled preprint study comparing single-agent and multi-agent configurations on the ALFWorld benchmark reports that a single evolved executor agent consumed an average of 5,400 tokens and 18.0 language-model calls per task, while a three-role team (planner, executor, critic) consumed 9,700 tokens and 32.2 calls per task, roughly 1.8 times the token cost and 1.8 times the call count [16]. The same study reports the team configuration achieved a higher mean success rate (0.769) than the single-agent configuration, but that the difference was not statistically distinguishable from the single-agent result ($p = 0.80$) at the sample sizes tested [16].

This is a single benchmark, a single preprint, and a specific multi-agent architecture (a fixed planner-executor-critic team compared against one evolved single agent); it should not be read as a general finding about all multi-agent systems. Within that scope, however, it is a materially cleaner result than a simple token-cost comparison would suggest: the reported evidence is that a substantially higher resource cost purchased no statistically distinguishable improvement in task success on this benchmark. Other research on multi-agent token and latency overhead more broadly finds that added resource cost is not uniformly justified across tasks and architectures, and that the literature is still in the process of separating cases where added agentic structure earns its cost from cases where it does not [16]. This paper does not claim multi-agent architectures are wasteful in general; it claims their resource cost is a genuine cost that the existing literature evaluates primarily against task performance, and that this paper's framework adds a second comparison the literature has not yet made.

7.2 Reframing the comparison

Section 6's two-ledger structure makes that second comparison explicit. Directing additional hardware and energy toward more agents and more inference (Strategy A) increases the primary-product resource footprint in Table 1 and, per Section 4, may accelerate the point at which a unit's AI-specific useful life ends. Directing the same constrained physical resources, including hardware already past its AI-specific useful life, toward the secondary uses catalogued in Table 2 (Strategy B) is the alternative this paper's framework makes visible as a genuine comparison rather than treating Strategy A's cost only against doing nothing with the resource.

The question this paper poses, and does not itself answer with a number, is: given a constrained hardware resource, at what point does the marginal value of directing an additional unit of that resource toward primary-product agentic compute fall below the marginal value available from directing the same unit toward the secondary economy in Table 2? This is a resource-allocation question, conditional on the specific task and deployment; it is not a claim that Strategy B is generally preferable to Strategy A, and answering it for any real deployment requires the fleet-level data this paper does not have (Section 8.2).

The bridge from Section 3's obsolescence argument to this application is direct: adding agentic structure to a pipeline is not simply "more compute." It is a decision to direct a unit of hardware capacity, with a finite AI-specific useful life (Section 3.1), toward marginal primary-product output rather than toward the secondary-economy value that same unit could otherwise generate once its AI-specific useful life is judged to have ended. Greater agentic utilization plausibly shortens the interval before a fleet operator must make that judgment for a given unit, though this paper's evidence does not establish the magnitude of that effect, only the structural mechanism.

8. Limitations

This section states the paper's limitations by category, following the taxonomy below, so that each limitation is attributable to a specific source rather than treated as a single general caveat. These are stated as boundaries of the present contribution, not as apologies for it.

8.1 Identification limitations

The depreciation-schedule divergence documented in Section 4.2 is consistent with, but does not by itself establish, workload-specific technological obsolescence as its cause. A longer assumed useful life reduces reported depreciation expense and improves near-term reported earnings; this paper cannot rule out that some part of the observed schedule changes reflects financial-reporting incentive rather than a genuine reassessment of economic service life. What partially supports treating the changes as at least in part genuine is that Amazon and Meta moved in opposite directions on comparable asset categories in the same window [9][10], a pattern that would be harder to produce through earnings management alone than a uniform industry-wide extension would be. This is suggestive, not dispositive, and the paper does not claim more than that.

8.2 Data and source-verification limitations

Several quantitative claims in this paper rest on industry reporting this research pass could not independently corroborate against a primary source, and are flagged here rather than presented with unqualified confidence: the reported 95 percent rebooking price for redeployed GPUs from expired contracts (Section 4.3); the 20 to 40 percent tariff-driven cost increase and the estimated 12.12 to 12.95 billion dollar IT asset disposition market size (Section 5.2); and the widely circulated 185 million ton CO2-equivalent annual figure for global integrated circuit production, which recurs across secondary industry sources without a traceable primary methodology (Section 5.3). None of these figures is load-bearing for the paper's structural argument, which depends on the qualitative pattern (displaced hardware retains value; embodied footprint is measurable at the component level) rather than on any single quantitative estimate, but each should be independently verified before being relied upon in a context where precision matters.

8.3 Measurement limitations

Effective Hardware Life Cost is introduced in Section 3.2 as a conceptual formulation. It has not been estimated for any real deployment in this paper, because doing so would require fleet-level utilization data, realized acquisition pricing, site-specific energy costs, and disposition outcomes that are not public at the granularity the formulation requires. The formulation states what would need to be measured; it does not constitute a measurement.

8.4 Allocation limitations

This paper argues that an allocation mechanism is needed to distribute an asset's embodied resource footprint across its primary and secondary uses (Section 6.4), and identifies five candidate bases, but does not select among them, apply one to real fleet data, or produce a numeric estimate of how much of any deployment's footprint should be reassigned to secondary use. The paper's contribution at this stage is establishing that the allocation question exists and is methodologically non-trivial, not resolving it.

8.5 Generalizability limitations

The multi-agent evidence in Section 7.1 is drawn from a single preprint study, a single benchmark (ALFWorld), and a specific architecture comparison. It should not be generalized to all multi-agent systems, all tasks, or all benchmarks. The paper's application of the two-ledger framework to multi-agent architecture (Section 7.2) is a structural reframing of the comparison researchers should make, not an empirical finding about multi-agent systems in general.

8.6 Accounting-interpretation limitations

Secondary-market prices (Section 4.3) demonstrate that displaced hardware retains economic value; they do not, by themselves, establish what the environmentally or resource-optimal allocation of embodied footprint across an asset's uses should be. Section 5.1 states this distinction explicitly because it is easy to conflate an empirical observation (value persists) with a normative conclusion (therefore this is the correct attribution rule). This paper takes the first as established by the evidence in Section 4.3 and treats the second as an open methodological question, not as something the first implies.

9. Discussion and Implications

For a firm evaluating its own AI infrastructure investment, the practical implication of the EHLC concept (Section 3.2) is that a lower headline price per unit of compute is not equivalent to a lower cost per unit of useful work delivered over the period a unit remains competitive for its intended workload. A firm choosing between higher-performance and lower-cost hardware may find the lower-cost option less competitive on cost per inference today while more competitive on value per unit of embodied resource once effective useful life is accounted for; a resource-adjusted ROI calculation that stops at Hardware Capital Cost without asking how long the acquired capacity remains competitive will not surface that trade-off.

For a firm or policymaker evaluating industry material-footprint claims, Sections 5 and 6 together imply that figures citing the full embodied carbon or water cost of AI hardware against only its first, frontier-AI use are likely to overstate that deployment's attributable share of the footprint, to the extent the same hardware subsequently generates value in other uses. As stated in Section 5.3 and repeated here because the distinction is easily lost: correcting this overstatement does not reduce

the absolute resource footprint already incurred in manufacturing the hardware. It changes which economic activity should be charged for which share of an already-fixed quantity, which matters for anyone comparing AI's resource intensity against alternative uses of the same capital, but it is not a claim that AI's true environmental footprint is smaller than reported.

For the specific question of multi-agent architecture, Section 7's reframing suggests that the relevant design and investment question is not only "does this multi-agent pipeline improve task performance enough to justify its added token cost," a question existing literature is already examining, but "does directing this hardware capacity toward additional agentic compute outperform directing the same capacity toward the secondary economy this paper describes." This paper's own evidence does not answer that second question for any real deployment; it argues the question is well-posed and states what would be needed to answer it.

10. Open Questions for Subsequent Research

This paper closes the open question left by the programme's first four papers concerning what happens to hardware capacity over time and what should be credited or charged as a result (Section 2). It opens the following for subsequent work, organized as a sequence rather than a flat list, since each later question depends on progress on the one before it:

1. **Allocation methodology.** Which of the five candidate allocation bases identified in Section 6.4, compute utilization, useful compute delivered, revenue generated, operating hours, or physical resource depletion, produces the most defensible attribution of embodied resource cost to a hardware unit's primary AI use, and does the answer differ across the four residual categories distinguished in Section 6.3?
2. **Empirical estimation.** Once an allocation basis is chosen, what does applying it to real or reconstructed fleet data imply for the resource-adjusted ROI of specific classes of AI deployment?
3. **The marginal-value comparison.** At what measured point, for a real fleet with real utilization data, does the marginal value of additional primary-product agentic compute fall below the marginal value available from directing the same resource toward the secondary economy described in Table 2 (Section 7.2)?
4. **Series integration.** Does the compression of AI-specific useful life documented in Section 4 change the economics of the Specialised Digital Asset hierarchy this programme's first paper proposed, and if so, in which direction [4]?
5. **Distinguishing the identification alternative.** Is there a measurable relationship between an operator's stated hardware depreciation schedule and the actual secondary-market disposition of its retired fleet, which would help distinguish Section 8.1's financial-engineering alternative explanation from this paper's workload-specific obsolescence account?

11. Conclusion

This paper's argument rests on two observations that pull in opposite directions and must be held together. A physical accelerator's economic competitiveness for the specific workload it was acquired for can decline well before the unit itself fails, which existing cost accounting, anchored to Hardware Capital Cost and a single useful-life figure, does not capture. At the same time, an accelerator displaced from that workload does not generally stop generating economic or material value; it commonly continues in a second, third, or further role before final material recovery, which a framework charging its entire embodied footprint against the first deployment also does not capture.

The paper's proposed response to both observations is the same structural move: track a hardware asset's lifecycle rather than only its first chapter. Effective Hardware Life Cost makes the first observation explicit as a conceptual complement to Hardware Capital Cost. The two-ledger structure makes the second observation explicit by keeping a primary-product ledger and a byproduct ledger separately visible rather than netting one against the other. Applied to the industry's direction toward more compute-intensive multi-agent architectures, this structural move reframes the question from whether additional agents are worth their token cost, which existing research is already examining, to whether the hardware capacity those agents consume would generate more total value directed toward the secondary economy instead, a comparison this paper's evidence does not resolve but argues is well-posed and currently absent from the literature.

What this paper establishes is the structural case for that comparison and the conceptual apparatus, EHLC, the five lifetimes, the two-ledger structure, needed to eventually make it. What it does not establish is a numeric answer to the allocation question in Section 6.4, an empirical resolution of the marginal-value comparison in Section 7.2, or a claim about how any specific firm or fleet should act. Those remain, as stated in Section 10, questions for the next stage of this research programme.

12. References

Sources are labeled by type where that materially affects how a claim should be weighted: [Primary] for regulatory filings and company disclosures, [Preprint] for non-peer-reviewed academic work, [Industry] for trade press and industry analysis.

1. Introl, "Secondary GPU Markets: Buying and Selling Used AI Hardware," 2025. [Industry] introl.com/blog/secondary-gpu-markets-buying-selling-used-hardware-guide-2025
2. GPUnex, "Buying or Selling GPUs in 2026: Prices, Tips & Rent vs Sell Analysis," 2026. [Industry] gpunex.com/blog/buying-selling-gpus-2026
3. Hashrate Index, "Used GPU Market: A100 & H100 Pricing, Depreciation," 2026; Big Data Supply, "Selling AI GPUs: The Secondary Market for A100s & H100s." [Industry] hashrateindex.com/blog/used-gpu-market-pricing-deprecation-secondary-ai ; bigdatasupply.com/selling-ai-gpu-the-secondary-market-for-a100s-and-h100s
4. triNetra Research, "External AI Dependence and Startup Financial Survivability," EAD-2026-01, July 2026.

5. triNetra Research, "The Judgment Layer: An Inductive Theory of Understanding Synthesis Failure in Large Enterprises," EAD-2026-02, July 2026.
6. triNetra Research, "The Infrastructure Loop: Digital Asset Ownership, Distributed Infrastructure Participation, and Economic Resilience," EAD-2026-03, July 2026.
7. SemiAnalysis, "Nvidia Blackwell Perf TCO Analysis: B100 vs B200 vs GB200NVL72," and Civo, "Comparing NVIDIA's B200 and H100," 2025-2026. [Industry] newsletter.semianalysis.com/p/nvidia-blackwell-perf-tco-analysis ; civo.com/blog/comparing-nvidia-b200-and-h100
8. SiliconANGLE, "Resetting GPU depreciation: Why AI factories bend, but don't break, useful life assumptions," November 2025. [Industry] siliconangle.com/2025/11/22/resetting-gpu-depreciation-ai-factories-bend-dont-break-useful-life-assumptions
9. National Law Review, citing Deep Quarry analysis of Amazon and Meta 2025 fiscal disclosures, "Artificial Intelligence GPU Depreciation Debate and Earnings Risk." [Industry, citing Primary filings] natlawreview.com/article/deep-quarry-useful-lives-gpus-key-considerations
10. Amazon.com, Inc., Form 10-Q for the quarterly period ended September 30, 2025, U.S. Securities and Exchange Commission EDGAR filing. [Primary] sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001018724&type=10-Q (nine-month useful-life change disclosure: increase in depreciation and amortization expense of $889 million, reduction in net income of $677 million, for the nine months ended September 30, 2025)
11. GPUnex, "Buying or Selling GPUs in 2026," depreciation-curve data. [Industry] gpunex.com/blog/buying-selling-gpus-2026
12. NVIDIA Corporation, 2025 Product Carbon Footprint disclosure for the HGX H100 system, as cited and used as a validation benchmark in: arXiv preprint, "More than Carbon: Cradle-to-Grave environmental impacts of GenAI training on the Nvidia A100 GPU," 2025. [Primary disclosure, cited via Preprint] arxiv.org/html/2509.00093v1
13. Interface EU, "Chip Production's Ecological Footprint: Mapping Climate and Environmental Impact." [Industry/policy research] interface-eu.org/publications/chip-productions-ecological-footprint
14. TSMC, 2022 Sustainability Report, water resource management disclosure (approximately 35 billion gallons / 132 billion liters of ultrapure water used globally in 2022, a 21 percent increase over 2021). [Primary] esg.tsmc.com/en/focus/greenManufacturing/waterResourceManagement.html
15. imec, "How can we reduce environmental impact in chip manufacturing?"; Semiconductor Digest, republished analysis. Figure (185 million tons CO2-equivalent annual IC production emissions) recurs across these and other secondary sources without a traceable primary methodology document identified in this research pass; flagged for author verification. [Industry, unverified primary source] imec-int.com/en/articles/how-can-we-reduce-environmental-impact-chip-manufacturing
16. Dylan, D., Brennan, A., Murphy, C., O'Sullivan, N., Kelly, C., and Walsh, S., "At Equal Inference Cost, Multi-Agent Structure Does Not Beat a Single Frozen Agent," arXiv preprint, 2026. [Preprint] arxiv.org/html/2609.04217