Computational Systems & Infrastructure
METHODS REVIEW | COMMISSIONED SYNTHESIS

Reproducible Performance Engineering for Heterogeneous Accelerators

James Anderson - Corresponding author

Abstract

This methods review examines reproducible performance engineering for heterogeneous accelerators. The organizing question is which experimental controls make performance comparisons portable across hardware and software stacks. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to presenting a narrowly tuned speedup as a generally reproducible systems result. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in high-performance and scientific computing evaluations.

Keywords: accelerators, GPU benchmarking, reproducibility, performance engineering, scientific computing

Synthesis lens	Principal question
Mechanism	the stack from algorithm and compiler through runtime, memory hierarchy, device, and measurement tooling
Evaluation	warm-up, variance, power limits, numerical fidelity, workload representativeness, and uncertainty
Primary risk	presenting a narrowly tuned speedup as a generally reproducible systems result

1. Introduction

Research on reproducible performance engineering for heterogeneous accelerators sits at the boundary between a promising component and a consequential decision. The core question is which experimental controls make performance comparisons portable across hardware and software stacks. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is presenting a narrowly tuned speedup as a generally reproducible systems result. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Abadal et al. (2021) addresses "Computing Graph Neural Networks: A Survey from Algorithms to Accelerators." In this synthesis, that work is used as a source on problem definition within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Pandey et al. (2022) addresses "The transformational role of GPU computing and deep learning in drug discovery." In this synthesis, that work is used as a source on conceptual boundary within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Shalf (2020) addresses "The future of computing beyond Moore’s Law." In this synthesis, that work is used as a source on measurement within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the stack from algorithm and compiler through runtime, memory hierarchy, device, and measurement tooling. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Bridges et al. (2016) addresses "Understanding GPU Power." In this synthesis, that work is used as a source on system mechanism within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Varghese et al. (2021) addresses "A Survey on Edge Performance Benchmarking." In this synthesis, that work is used as a source on representation choice within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Ye et al. (2023) addresses "Deep Learning Workload Scheduling in GPU Datacenters: A Survey." In this synthesis, that work is used as a source on failure analysis within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on warm-up, variance, power limits, numerical fidelity, workload representativeness, and uncertainty. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Deelman et al. (2017) addresses "The future of scientific workflows." In this synthesis, that work is used as a source on comparative evaluation within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Humble et al. (2021) addresses "Quantum Computers for High-Performance Computing." In this synthesis, that work is used as a source on uncertainty within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to high-performance and scientific computing evaluations depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is versioned environments, executable artifacts, independent reruns, and disclosure of tuning budgets. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Caviedes-Voullième et al. (2023) addresses "SERGHEI (SERGHEI-SWE) v1.0: a performance-portable high-performance parallel-computing shallow-water solver for hydrology and environmental hydraulics." In this synthesis, that work is used as a source on implementation within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Silva et al. (2023) addresses "A review on the decarbonization of high-performance computing centers." In this synthesis, that work is used as a source on governance within reproducible performance engineering for heterogeneous accelerators. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for high-performance and scientific computing evaluations. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For reproducible performance engineering for heterogeneous accelerators, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to reproducible performance engineering for heterogeneous accelerators: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in high-performance and scientific computing evaluations requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

References

Abadal, S., Jain, A., Guirado, R., López-Alonso, J., & Alarcón, E. (2021). Computing Graph Neural Networks: A Survey from Algorithms to Accelerators. ACM Computing Surveys, 54(9), 1-38. https://doi.org/10.1145/3477141

Bridges, R. A., Imam, N., & Mintz, T. M. (2016). Understanding GPU Power. ACM Computing Surveys, 49(3), 1-27. https://doi.org/10.1145/2962131

Caviedes-Voullième, D., Morales-Hernández, M., Norman, M. R., & Özgen-Xian, I. (2023). SERGHEI (SERGHEI-SWE) v1.0: a performance-portable high-performance parallel-computing shallow-water solver for hydrology and environmental hydraulics. Geoscientific model development, 16(3), 977-1008. https://doi.org/10.5194/gmd-16-977-2023

Deelman, E., Peterka, T., Altintas, I., Carothers, C. D., Dam, K. K. V., Moreland, K., Parashar, M., Ramakrishnan, L., Taufer, M., & Vetter, J. (2017). The future of scientific workflows. The International Journal of High Performance Computing Applications, 32(1), 159-175. https://doi.org/10.1177/1094342017704893

Humble, T. S., McCaskey, A., Lyakh, D. I., Gowrishankar, M., Frisch, A., & Monz, T. (2021). Quantum Computers for High-Performance Computing. IEEE Micro, 41(5), 15-23. https://doi.org/10.1109/mm.2021.3099140

Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A. C., & Cherkasov, A. (2022). The transformational role of GPU computing and deep learning in drug discovery. Nature Machine Intelligence, 4(3), 211-221. https://doi.org/10.1038/s42256-022-00463-x

Shalf, J. (2020). The future of computing beyond Moore’s Law. Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences, 378(2166), 20190061. https://doi.org/10.1098/rsta.2019.0061

Silva, C. A., Vilaça, R., Pereira, A., & Bessa, R. J. (2023). A review on the decarbonization of high-performance computing centers. Renewable and Sustainable Energy Reviews, 189, 114019. https://doi.org/10.1016/j.rser.2023.114019

Varghese, B., Wang, N., Bermbach, D., Hong, C. H., Lara, E. D., Shi, W., & Stewart, C. (2021). A Survey on Edge Performance Benchmarking. ACM Computing Surveys, 54(3), 1-33. https://doi.org/10.1145/3444692

Ye, Z., Gao, W., Hu, Q., Sun, P., Wang, X., Luo, Y., Zhang, T., & Wen, Y. (2023). Deep Learning Workload Scheduling in GPU Datacenters: A Survey. ACM Computing Surveys, 56(6), 1-38. https://doi.org/10.1145/3638757