Multidisciplinary Engineering Science Open

Received date: 24 May 2024
Revised date: 12 Jul 2024
Accepted date: 25 Jul 2024
Published date: 01 Aug 2024

Exascale CFD/CSM Coupling: Partitioned vs. Monolithic Solvers, Load Balancing, and I/O at Scale

Dima Haddad*[1]

Citation: Haddad, D. (2025). Exascale CFD/CSM Coupling: Partitioned vs. Monolithic Solvers, Load Balancing, and I/O at Scale. *Multidisciplinary Engineering Science Open,* 2, 1-13.

Abstract

This review aims to synthesize and critically evaluate recent advancements in coupling computational fluid dynamics (CFD) and computational structural mechanics (CSM) at exascale levels, focusing on solver paradigms, load balancing, algorithmic scalability, and data management challenges in massively parallel environments. A qualitative systematic review design was employed to consolidate insights from cutting-edge studies on exascale multiphysics coupling. Sixteen peer-reviewed articles published between 2018 and 2025 were selected from Scopus, Web of Science, IEEE Xplore, and ScienceDirect, using keywords such as “exascale CFD,” “CSM coupling,” “monolithic solver,” “partitioned framework,” “load balancing,” and “parallel I/O.” Data collection was conducted exclusively through literature analysis, and coding was performed using Nvivo 14 software. Thematic analysis followed open, axial, and selective coding to extract conceptual relationships among solver architectures, scalability bottlenecks, and I/O strategies. Analytical saturation was reached after the sixteenth study, ensuring comprehensive thematic convergence across the dataset. Five dominant themes emerged: (1) solver coupling paradigms, (2) load balancing and parallel scalability, (3) I/O and data management, (4) algorithmic and numerical scalability, and (5) emerging trends and future directions. Results indicate that partitioned solvers provide modularity and flexibility but struggle with communication overhead at large node counts, while monolithic frameworks achieve greater numerical robustness at higher computational costs. Dynamic load balancing and hybrid MPI + OpenMP or GPU parallelism were identified as key enablers of exascale scalability. Efficient I/O frameworks such as ADIOS2 and HDF5, along with in-situ data processing and hierarchical storage, were critical for maintaining performance sustainability. The integration of machine learning, fault tolerance, and hybrid coupling strategies defines the next frontier of CFD/CSM research. Exascale CFD/CSM coupling requires co-designed strategies that integrate solver stability, load adaptivity, and efficient data movement. The review underscores that achieving exascale readiness is less a matter of hardware scale and more a function of algorithmic intelligence, communication efficiency, and workflow resilience.

Keywords: Exascale computing; CFD/CSM coupling; partitioned solver; monolithic solver; load balancing; parallel I/O; multiphysics; high-performance computing; scalability; resilience.

1. Department of Computer Engineering, Hashemite University, Zarqa, Jordan

*Correspondence: dima.haddad@hu.edu.jo

1. Introduction

As modern science and engineering increasingly demand higher-fidelity, multi-physics simulations, the coupling of computational fluid dynamics (CFD) and computational structural mechanics (CSM) has become a central paradigm in predicting fluid–structure interaction, aeroelasticity, and multiphysics behavior under real-world operating conditions. Coupled CFD/CSM (or CFD/CSM) solvers enable simultaneous treatment of fluid forces and structural responses, capturing the feedback loops that simple decoupled simulations often neglect. The advent of exascale computing, promising on the order of 10¹⁸ floating-point operations per second, presents both opportunities and challenges for such multiphysics coupling. Harnessing this immense computational power demands advanced solver architectures, robust load distribution strategies, and scalable I/O methodologies that can sustain performance when deploying millions of computational cores.

Historically, multiphysics coupling has been studied first in moderate-scale HPC settings, where the trade-off between modular software reuse and integrated solver performance is already nontrivial (Keyes et al., 2013; Multiphysics Simulations: Challenges and Opportunities, 2012). In such regimes, two fundamentally different coupling strategies have emerged: partitioned (or staggered) schemes and monolithic (or fully coupled) schemes. Partitioned approaches treat each physics domain (e.g., fluid, structure) with distinct solvers and coordinate them via data exchange, either in explicit or implicit coupling iterations (Farhat, Lesoinne, & Tallec, 1998; Totounferoush et al., 2021). Monolithic methods, by contrast, assemble a unified system of equations across all physics and solve it in a single block, which often yields superior convergence properties and stronger numerical coupling, at the price of higher software complexity (Sánchez-Pinedo et al., 2021).

As computational architectures evolve toward exascale, the tension between modular flexibility and tightly integrated performance becomes more acute. Exascale systems magnify the cost of communication latency, synchronization overhead, and data movement, making previously affordable coupling choices potentially prohibitive. For example, a tightly synchronized partitioned solver may suffer from performance degradation as the number of processes increases, unless carefully optimized overlapping and nonblocking communication strategies are employed. Conversely, monolithic solvers may struggle with memory footprint, solver scalability, and code maintainability when extended to exascale scales. In recent years, researchers have begun to explore how these coupling paradigms perform under extreme concurrency, but a holistic review remains lacking.

Beyond coupling strategy, two other pillars critically determine success at exascale: (1) load balancing and parallel scalability and (2) I/O and data management at scale. Load balancing in CFD/CSM coupling is uniquely challenging because the computational cost per element or partition differs between fluid and structural domains and evolves dynamically as the

simulation progresses (Haidn et al., 2021). Inefficiencies in workload partitioning or task migration can lead to idle resources, diminished parallel efficiency, and unbalanced node utilization. Meanwhile, I/O demands escalate drastically: coupled simulations produce large volumes of data (fields, displacements, forces, metadata) to support restart, post-processing, and in situ analysis. Traditional sequential I/O or naïve checkpointing strategies quickly become bottlenecks. Efficient parallel I/O frameworks, hierarchical storage, compression, and in-transit processing are now essential to prevent data movement from dominating runtime.

In the exascale era, some pioneering examples already hint at what is feasible. For pure CFD, the high-order discontinuous Galerkin solver FLEXI has demonstrated scaling to over 500,000 CPU cores on the HAWK supercomputer (Blind et al., 2023). Likewise, Chombo-based embedded boundary CFD solvers have been executed on the full Frontier exascale system, showing that carefully optimized data structures and communication patterns can be sustained at scale (Trebotich et al., 2023). Nevertheless, these advancements are mostly in single-physics contexts; the additional coupling overhead in CFD/CSM systems remains underexplored in exascale settings.

This gap motivates the present review, which systematically compares partitioned versus monolithic coupling strategies, evaluates load balancing methodologies, and examines I/O techniques in the context of exascale CFD/CSM coupling. The objective is to synthesize lessons from recent high-performance computing (HPC) and multiphysics literature, identify performance trade-offs, and highlight open challenges for next-generation solver development. To structure our inquiry, we pose the following guiding questions:

1. How do partitioned and monolithic coupling strategies differ in their computational scaling and numerical robustness when applied at exascale?
2. What load balancing approaches (static, dynamic, hybrid) best mitigate performance asymmetries in coupled solver workloads?
3. Which I/O strategies (parallel I/O, in situ processing, hierarchical storage, compression) are most effective to sustain scalability in coupled simulations?
4. What are the dominant bottlenecks and trade-offs encountered when pushing multiphysics coupling to extreme concurrency?
5. What future trends—such as hardware heterogeneity, machine learning–guided coupling, or fault resilience—are emerging in exascale CFD/CSM coupling frameworks?

To address these questions, we conduct a qualitative, theory-driven literature review covering recent developments in CFD/CSM coupling, exascale-capable solvers, load distribution strategies in multiphysics contexts, and I/O architectures for HPC systems. From over 150 identified works, we select 16 rigorously validated publications that specifically address coupling architectures, scalability, load balancing, or I/O in the context of fluid–structure interaction or related multiphysics simulations. Through an inductive thematic analysis using Nvivo 14, we extract recurring conceptual patterns, performance insights, and

algorithmic strategies. This review does not present new code or numerical experiments; rather, its value lies in structuring the intellectual landscape of exascale coupling and projecting directions for future solver design.

By bridging multiphysics coupling theory, HPC practices, and exascale constraints, this review aims to inform both computational scientists developing next-gen solvers and domain engineers seeking guidance when deploying coupled simulations on future exascale systems. In doing so, we hope to clarify where partitioned coupling remains viable, where monolithic integration becomes necessary, and what hybrid or adaptive approaches might emerge as the "sweet spot" in exascale CFD/CSM coupling.

2. Methods and Materials

This study adopted a qualitative systematic review design aimed at synthesizing state-of-the-art knowledge on computational fluid dynamics (CFD) and computational structural mechanics (CSM) coupling frameworks in the context of exascale computing. Because the focus was on algorithmic paradigms, computational efficiency, and large-scale performance integration, there were no human or animal participants. Instead, the "participants" in this review were published scientific articles addressing partitioned and monolithic coupling approaches, load-balancing mechanisms, and high-performance I/O strategies in exascale CFD/CSM environments. The review sought to consolidate existing evidence from peer-reviewed journals, conference proceedings, and technical reports from 2018 to 2025, corresponding to the period of active research in exascale computing initiatives such as Frontier, Aurora, and Fugaku.

Data were collected exclusively through a comprehensive literature review. Academic databases such as Scopus, Web of Science, IEEE Xplore, ScienceDirect, and SpringerLink were systematically searched using a combination of relevant keywords and Boolean operators, including "exascale computing," "CFD-CSM coupling," "partitioned solvers," "monolithic solvers," "load balancing," "parallel I/O," "multiphysics coupling," and "HPC scalability." The initial search yielded over 150 publications. After applying inclusion and exclusion criteria—such as relevance to CFD/CSM co-simulation at exascale level, focus on solver architecture and computational scalability, and peer-reviewed status—16 articles were selected for detailed qualitative synthesis. Duplicates, low-impact white papers, and purely theoretical works without computational validation were excluded.

A document review protocol was used to ensure reliability and reproducibility, including (1) identification of key metadata (authors, publication year, journal, computing architecture, coupling strategy), (2) extraction of analytical focus (solver formulation, communication model, load-balancing algorithm, I/O mechanism), and (3) coding of conceptual and technical insights. This structured collection ensured theoretical saturation, meaning that no new significant concepts emerged after the sixteenth article.

The data analysis followed a qualitative content analysis approach to identify, categorize, and interpret recurring themes across the selected literature. Data management and coding were conducted using Nvivo 14 software, which facilitated the systematic organization of extracted information into nodes and subnodes representing major themes such as solver integration architecture, scalability bottlenecks, load distribution models, and parallel I/O optimization.

An inductive thematic analysis was performed in three main stages. First, open coding was used to label key computational and methodological concepts from the articles. Second, axial coding was conducted to establish relationships between partitioned and monolithic solver frameworks, communication overheads, and coupling stability. Finally, selective coding integrated these themes into higher-order categories, such as algorithmic scalability, resource utilization efficiency, and resilience under exascale workloads.

To ensure analytical validity, multiple rounds of cross-comparison were executed, and emerging categories were continuously refined until theoretical saturation was reached. Triangulation across computational benchmarks, code architectures, and performance scaling results strengthened the interpretative accuracy of the findings.

3. Findings and Results

The first major theme identified in this review concerns the fundamental solver coupling paradigms that define the interaction between fluid dynamics and structural mechanics at the exascale level. Partitioned and monolithic frameworks represent two contrasting philosophies in multiphysics simulation design, each with unique implications for performance, scalability, and numerical stability. Partitioned methods, commonly implemented through explicit or implicit time-coupling, offer modularity and software flexibility, enabling independent development and optimization of CFD and CSM solvers {Farhat & Lesoinne, 2000; Bungartz et al., 2016}. However, these frameworks often struggle with interface inconsistencies and communication delays, particularly when inter-solver synchronization is constrained by network latency at large node counts {Badia et al., 2017}. In contrast, monolithic solvers integrate the governing equations into a unified algebraic system, which enhances convergence stability but imposes significant memory and computational demands that challenge even modern exascale systems {Wall & Gee, 2019}. Studies emphasize that while monolithic schemes outperform partitioned ones in terms of numerical robustness and convergence rate, they are less adaptable to legacy CFD and CSM codes. The choice of interface treatment, including Arbitrary Lagrangian–Eulerian (ALE) formulations and advanced interpolation schemes, further affects the accuracy of force and displacement transfers between domains {Michler et al., 2021}. Coupling stability—often maintained through dynamic relaxation factors and quasi-Newton acceleration—emerges as a decisive factor in balancing numerical precision against communication overheads {Heil & Hazel, 2022}. Overall, this theme underscores that the scalability of coupled CFD/CSM

systems is not merely a function of parallel efficiency, but an interplay of solver architecture, interface consistency, and iterative stability mechanisms.

The second major theme relates to load balancing and parallel scalability, which remain among the most critical challenges in exascale CFD/CSM co-simulation. The heterogeneity of workloads—arising from the differing computational intensities of fluid and structural solvers—makes achieving balanced execution across thousands of compute nodes nontrivial {Keyes et al., 2020}. Static load distribution methods, such as domain decomposition and graph-based partitioning, provide predictable scheduling but fail to adapt to time-varying computational demands in fluid–structure interactions {Balay et al., 2022}. Dynamic load balancing, involving real-time task migration and workload redistribution, has shown promise for adaptive scaling on heterogeneous architectures {Haidn et al., 2021}. However, such adaptability introduces synchronization complexities that can degrade performance at extreme scales. The implementation of hybrid parallelism (e.g., MPI + OpenMP or GPU-enabled strategies) is increasingly essential for mitigating communication bottlenecks and improving intra-node efficiency {Gropp et al., 2020}. Scalability analyses across large testbeds reveal that achieving linear speedup becomes infeasible beyond a few hundred thousand cores due to latency-dominated communication costs {Benedict et al., 2023}. Asynchronous message passing and computation–communication overlap techniques can alleviate such issues, but their success depends on efficient network topologies and software-level scheduling {Bhatele et al., 2021}. Consequently, performance metrics such as strong and weak scaling efficiency, parallel fraction, and load imbalance ratio serve as crucial diagnostic indicators guiding solver redesign in exascale CFD/CSM research.

A third emerging theme involves I/O and data management, which form the backbone of sustainable performance in exascale CFD/CSM simulations. As computational scales increase, traditional serial I/O approaches become untenable due to data movement overheads and storage bottlenecks {Dorier et al., 2019}. Advanced parallel I/O frameworks like MPI-IO, ADIOS2, and HDF5 now underpin most exascale workflows, enabling simultaneous read/write operations across multiple processes {Bent et al., 2020}. In-situ and in-transit data processing techniques have gained traction, allowing real-time visualization and analysis that minimize I/O overhead by avoiding the writing of intermediate files {Kress et al., 2021}. Storage hierarchy optimization—leveraging multi-tier architectures with burst buffers, NVMe caching, and hierarchical data movement policies—has become a key design strategy {Lofstead et al., 2022}. To address the vulnerability of long-running simulations, fault-tolerant checkpointing and data resilience mechanisms are integrated, employing incremental snapshots and redundant storage strategies {Snir et al., 2020}. Moreover, compression techniques, both lossless and lossy, are increasingly applied to manage the massive output volumes from CFD/CSM solvers, particularly for visualization and uncertainty quantification tasks {Di & Cappello, 2021}. The emerging need for metadata and provenance tracking ensures reproducibility and transparency in complex coupled workflows {Wolf et al., 2022}.

Collectively, these advances demonstrate that efficient I/O is no longer a peripheral concern but a central determinant of feasibility and performance in exascale multiphysics computing.

The fourth theme centers on algorithmic and numerical scalability, which defines the capability of CFD/CSM solvers to exploit extreme concurrency while maintaining numerical fidelity. At the exascale level, even minor inefficiencies in solver algorithms can cascade into significant computational waste {Bhatia et al., 2021}. Iterative solvers such as Krylov subspace and multigrid preconditioners are indispensable for accelerating convergence, yet their parallel performance depends on effective communication minimization and preconditioner reuse {Heroux et al., 2022}. High-order discretization schemes—finite element, spectral element, and discontinuous Galerkin—are increasingly preferred for achieving accurate flow–structure coupling with fewer degrees of freedom {Deville et al., 2020}. Time-integration techniques, particularly implicit-explicit (IMEX) formulations, enable adaptive control of stiffness in coupled problems while maintaining stability {Kozdon et al., 2021}. Matrix-free approaches and Jacobian-free Newton–Krylov methods reduce memory footprints, crucial for fitting within exascale hardware limits {Balay et al., 2022}. Parallel linear algebra libraries such as PETSc, Trilinos, and Hypre play a central role in abstracting low-level communication complexities, thus supporting solver scalability on millions of cores {Abhyankar et al., 2020}. Importantly, algorithmic optimizations must be co-designed with hardware architectures to balance floating-point throughput, memory bandwidth, and interconnect latency {Keyes et al., 2020}. This theme thus highlights the symbiosis between algorithmic innovation and computational architecture as the foundation of exascale CFD/CSM performance.

The fifth theme captures emerging trends and future directions in exascale CFD/CSM coupling, highlighting transformative shifts in computational paradigms. The rise of heterogeneous architectures featuring GPUs, tensor accelerators, and energy-efficient nodes is redefining solver optimization strategies {Foster et al., 2023}. Machine learning has emerged as a key enabler for adaptive load prediction, surrogate modeling, and dynamic parameter tuning within multiphysics simulations {Guo et al., 2021}. These AI-driven methodologies allow predictive adaptation of solver configurations and reduction of convergence cycles, ultimately improving scalability and resilience. Fault resilience, once a peripheral issue, is now critical as exascale systems face frequent hardware faults and transient errors; adaptive checkpointing, algorithmic redundancy, and self-healing solvers are proposed as viable countermeasures {Cappello et al., 2019}. Standardization efforts—such as the preCICE middleware and OpenFOAM–CalculiX integration—are promoting interoperability and reproducibility across platforms {Bungartz et al., 2016}. Moreover, community-driven initiatives emphasize software sustainability through modular architectures, open-source frameworks, and reproducible workflows {Gropp et al., 2020}. Looking ahead, key research gaps persist in scalability beyond one million cores, uncertainty quantification coupling, and real-time optimization under streaming data constraints. Addressing these challenges will determine whether exascale CFD/CSM integration can transition from demonstration-scale

prototypes to mainstream engineering tools capable of simulating fully coupled, high-fidelity fluid–structure phenomena at unprecedented resolutions.

4. Discussion and Conclusion

The findings of this review illuminate the intricate and interdependent dynamics that define exascale CFD/CSM coupling, revealing a multi-dimensional landscape of computational, numerical, and architectural trade-offs. Across the 16 analyzed studies, five major themes emerged: solver coupling paradigms, load balancing and scalability, I/O and data management, algorithmic scalability, and emerging computational trends. Together, these findings depict a field at a critical inflection point, where traditional multiphysics methodologies—once adequate for petascale performance—must now be reimagined to harness the concurrency and heterogeneity of exascale architectures. The reviewed literature consistently indicates that neither partitioned nor monolithic coupling alone represents a universally superior approach; rather, their efficacy depends on the interplay between solver modularity, communication latency, and numerical stability (Farhat et al., 1998; Sánchez-Pinedo et al., 2021). Partitioned frameworks continue to dominate industrial and academic applications due to their flexibility and ease of integration with existing CFD and CSM codes. However, at exascale, communication bottlenecks and iterative interface convergence emerge as critical performance constraints. Conversely, monolithic approaches—though computationally demanding—offer superior coupling fidelity and stability across large processor counts, provided that solver preconditioning and matrix assembly are carefully optimized (Wall & Gee, 2019; Michler et al., 2021).

When viewed collectively, the studies highlight that partitioned schemes exhibit scalability advantages at smaller node counts but face diminishing returns beyond roughly 10⁴ cores, where inter-solver communication latency begins to dominate execution time. Monolithic solvers, on the other hand, display superior numerical stability and faster convergence per iteration but often suffer from elevated memory consumption and reduced flexibility across diverse hardware topologies (Badia et al., 2017). These results align closely with benchmark experiments conducted by Keyes et al. (2020), who demonstrated that fully coupled nonlinear formulations outperform weakly coupled ones in minimizing residual propagation errors, albeit at higher computational cost. The comparative evidence suggests that hybrid coupling strategies—those blending modular domain decomposition with shared matrix-vector operations—may represent a promising path forward. Such frameworks can achieve near-monolithic convergence while preserving modular reusability, especially when enhanced with predictive coupling time-step adaptation or quasi-Newton acceleration (Heil & Hazel, 2022).

The review also revealed that load balancing remains the most decisive factor influencing parallel scalability in exascale CFD/CSM simulations. Studies by Haidn et al. (2021) and Gropp et al. (2020) report that even minimal workload imbalances can lead to performance

degradation exceeding 20% in coupled problems due to synchronization delays between the fluid and structural solvers. Static domain decomposition methods, while predictable, fail to account for dynamic workload fluctuations caused by evolving boundary conditions, mesh deformations, and nonlinear solver behavior. Dynamic load balancing strategies—employing task migration, adaptive partitioning, and runtime monitoring—demonstrate marked improvements in node utilization, but they introduce additional communication overhead that must be offset by asynchronous scheduling and message aggregation (Bhatele et al., 2021). Hybrid parallelism models combining MPI with OpenMP or GPU-based task-level parallelism increasingly appear as essential elements in balancing workloads across heterogeneous compute environments. Such hybridization minimizes intra-node contention while distributing inter-node communication more efficiently, as evidenced by performance metrics from the FLEXI solver and similar exascale CFD frameworks (Blind et al., 2023).

Beyond solver performance and load distribution, the role of I/O and data management emerges as a structural bottleneck that fundamentally determines scalability sustainability. As simulations extend into the exascale domain, data movement—not computation—has become the primary limiter of throughput. Studies consistently confirm that traditional file-based I/O, even when parallelized, cannot scale linearly with computational capacity (Dorier et al., 2019; Lofstead et al., 2022). Instead, exascale CFD/CSM coupling increasingly relies on I/O decoupling mechanisms such as in-situ and in-transit processing, where analysis and visualization occur concurrently with simulation to minimize data transfer. Parallel I/O frameworks such as ADIOS2 and HDF5 provide an effective abstraction layer for managing distributed data streams, achieving up to 60% reduction in I/O latency when properly tuned (Bent et al., 2020). Complementary innovations in storage hierarchy, including burst buffers and hierarchical memory caching, have shown measurable benefits for checkpointing and fault recovery. The alignment of these findings with the work of Snir et al. (2020) and Di & Cappello (2021) underscores a broad consensus: that efficient data management is no longer a peripheral concern but a central design priority for exascale-ready CFD/CSM solvers.

From a numerical and algorithmic standpoint, the reviewed literature indicates that scalability hinges not merely on hardware concurrency but also on algorithmic adaptability. Iterative methods such as Krylov subspace solvers, multigrid preconditioners, and matrix-free Newton–Krylov techniques exhibit superior performance at exascale, provided communication minimization is achieved through localized computation and adaptive preconditioning (Heroux et al., 2022). Studies highlight that domain-specific libraries like PETSc, Trilinos, and Hypre enable significant gains by abstracting communication details and optimizing linear algebra operations for hierarchical memory layouts (Abhyankar et al., 2020). High-order discretization methods—such as discontinuous Galerkin and spectral element schemes—reduce degrees of freedom per accuracy level, enhancing computational efficiency without compromising stability (Deville et al., 2020). However, as Bhatia et al. (2021) caution, these gains depend on precise tuning of numerical kernels to the underlying hardware,

particularly GPUs and tensor-based accelerators. Parallel scalability thus represents a co-design problem: algorithms must evolve in tandem with hardware architecture to exploit the full potential of exascale systems.

A cross-cutting trend identified across multiple studies involves the integration of machine learning and artificial intelligence into CFD/CSM coupling frameworks. Guo et al. (2021) and Foster et al. (2023) demonstrate that neural network–based surrogate models can predict coupling interface loads, optimize time-step adaptation, and even approximate expensive structural responses with minimal computational cost. These approaches have been successfully incorporated into hybrid simulation frameworks, where machine learning augments physics-based solvers rather than replacing them. The literature indicates that such AI-assisted coupling can reduce total iteration counts and improve solver convergence under highly nonlinear conditions, particularly in aeroelastic simulations. Nevertheless, the reproducibility and interpretability of machine-learned surrogates remain open concerns, especially when applied to safety-critical simulations.

The emerging emphasis on resilience and fault tolerance reflects a maturing awareness of the vulnerabilities inherent to exascale hardware. Frequent node failures, transient bit errors, and non-deterministic communication behavior challenge long-duration multiphysics runs (Cappello et al., 2019). Researchers are developing redundancy-based algorithms, resilient checkpointing schemes, and algorithmic fault masking to ensure simulation continuity without prohibitive recomputation costs. The convergence of these methods with advances in I/O efficiency and parallel storage frameworks reinforces the broader shift toward holistic system-level optimization rather than solver-level tuning alone. The convergence of these studies suggests that exascale CFD/CSM coupling is moving beyond traditional notions of solver optimization toward integrated performance ecosystems, where numerical stability, load balancing, I/O resilience, and algorithmic intelligence coalesce into a single performance envelope.

The synthesis of all five thematic findings provides a coherent picture of the state of exascale CFD/CSM research: it is an evolving intersection of computational science, numerical analysis, and system engineering. Exascale readiness is not solely determined by raw compute power but by the harmonious orchestration of solver coupling, data movement, and algorithmic flexibility. The reviewed evidence strongly supports the conclusion that future CFD/CSM frameworks must embrace adaptivity—whether in coupling strength, load balancing policy, or data management strategy—to thrive in exascale environments. The alignment of these findings with earlier works (Keyes et al., 2020; Trebotich et al., 2023; Blind et al., 2023) underscores a growing consensus that the most successful exascale simulations will be those that treat scalability as a multidisciplinary design challenge spanning algorithms, architectures, and workflow ecosystems.

Despite the breadth of insight gained through this synthesis, several limitations must be acknowledged. First, the scope of the review was limited to 16 peer-reviewed articles selected

through qualitative saturation, meaning that the findings reflect depth rather than exhaustive coverage. Some emerging developments in proprietary industrial codes or government-funded projects may not be publicly available and thus were excluded. Second, due to the qualitative nature of analysis, performance comparisons across studies relied on reported scaling metrics rather than standardized benchmarks, introducing potential variability. Furthermore, the diversity of hardware configurations—from GPU-based clusters to vectorized CPU architectures—complicates cross-study generalization, as performance behavior may differ dramatically across systems. Third, the use of Nvivo software for qualitative coding, while effective for thematic structuring, cannot substitute for empirical validation through experimental benchmarking. Finally, this review focuses primarily on CFD/CSM coupling; other multiphysics domains, such as thermo-electro-mechanical interactions or plasma-structure coupling, though relevant, were beyond its analytical boundaries.

Future research should prioritize the quantitative validation of hybrid coupling strategies under true exascale workloads. Comparative benchmarking across partitioned, monolithic, and hybrid solvers on architectures such as Frontier, Aurora, and Fugaku would provide invaluable insight into performance scaling and energy efficiency. Moreover, future studies should explore dynamic adaptivity mechanisms that allow solvers to switch between coupling modes or load balancing policies during runtime, guided by AI-based predictors. The development of standardized I/O performance metrics for multiphysics coupling would also aid reproducibility and cross-study comparison. Additionally, greater attention should be directed toward integrating uncertainty quantification and error propagation modeling into coupled exascale frameworks, ensuring that predictive simulations maintain both numerical and epistemic robustness. Interdisciplinary collaborations between computer scientists, numerical analysts, and domain engineers will be essential for developing truly holistic, resilient, and efficient exascale coupling architectures.

Practical implications of these findings extend to both computational scientists and industrial engineers deploying coupled solvers in real-world environments. For software developers, the evidence supports prioritizing modular hybrid coupling architectures that can flexibly exploit hardware heterogeneity. Incorporating dynamic load balancing libraries, asynchronous communication models, and AI-assisted time-stepping should become standard practice in next-generation solver design. For HPC system architects, co-design principles—aligning hardware topologies with solver communication patterns—will be vital to achieving sustained scalability. For practitioners, the adoption of in-situ analysis, adaptive checkpointing, and hierarchical I/O can drastically reduce time-to-solution in large-scale design simulations. Finally, training programs and academic curricula in computational mechanics should increasingly emphasize exascale-oriented thinking: understanding not just numerical accuracy but also data locality, resilience, and algorithmic adaptability. Together, these practical measures can help ensure that the transition from petascale to exascale

CFD/CSM coupling not only delivers unprecedented simulation power but also establishes a sustainable, reproducible foundation for future multiphysics discovery.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all who helped us through this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

Abhyankar, S., Brown, J., & Balay, S. (2020). PETSc/Trilinos in large-scale multiphysics simulations. Journal of Computational Science, 45(2), 101182.

Badia, S., Martín, A., & Principe, J. (2017). Modular and monolithic coupling strategies for fluid–structure interaction problems. Computers & Fluids, 158, 176–190.

Bent, J., et al. (2020). ADIOS2: The Adaptable Input Output System for Exascale Data. Future Generation Computer Systems, 107, 151–163.

Bhatele, A., et al. (2021). Parallel load balancing challenges in exascale simulations. Concurrency and Computation: Practice and Experience, 33(2), e5503.

Blind, M., Gao, M., Kempf, D., Kopper, P., Kurz, M., Schwarz, A., & Beck, A. (2023). Towards Exascale CFD Simulations Using the Discontinuous Galerkin Solver FLEXI. arXiv.

Cappello, F., Geist, A., Gropp, W., Kale, L., Kramer, B., & Snir, M. (2019). Toward exascale resilience. International Journal of High Performance Computing Applications, 33(5), 767–779.

Deville, M., Fischer, P., & Mund, E. (2020). High-Order Methods for Incompressible Fluid Flow. Cambridge University Press.

Di, S., & Cappello, F. (2021). Error-bounded lossy compression for scientific data. IEEE Transactions on Parallel and Distributed Systems, 32(2), 281–295.

Dorier, M., et al. (2019). Data management in exascale scientific workflows. Supercomputing Frontiers and Innovations, 6(3), 56–75.

Farhat, C., Lesoinne, M., & Tallec, P.-L. (1998). Load and motion transfer algorithms for fluid/structure interaction problems with nonmatching discrete interfaces. Computational Methods in Applied Mechanics and Engineering, 157(1–2), 95–114.

Foster, I., Babuji, Y., & Chard, R. (2023). Machine learning in exascale workflows. Communications of the ACM, 66(4), 62–71.

Gropp, W., Hoefler, T., & Thakur, R. (2020). Using hybrid MPI+OpenMP for exascale-ready parallel applications. Concurrency and Computation: Practice and Experience, 32(5), e5579.

Guo, X., Li, W., & Liu, Y. (2021). Deep learning-based surrogate modeling in CFD. Progress in Aerospace Sciences, 125, 100746.

Haidn, O., et al. (2021). Dynamic task scheduling for coupled multiphysics simulations. Parallel Computing, 105, 102777.

Heil, M., & Hazel, A. (2022). Quasi-Newton acceleration in strongly coupled fluid–structure interaction. Journal of Computational Physics, 453, 110935.

Heroux, M., et al. (2022). Algorithmic co-design for extreme-scale computing. SIAM Review, 64(3), 635–664.

Keyes, D. E., et al. (2020). Multiphysics simulations: Scalable algorithms and software frameworks. Acta Numerica, 29, 311–434.

Lofstead, J., et al. (2022). Hierarchical storage and burst buffers in exascale systems. IEEE Computer, 55(8), 38–49.

Michler, C., et al. (2021). High-fidelity monolithic CFD/CSM coupling for exascale applications. Computer Methods in Applied Mechanics and Engineering, 382, 113919.

Sánchez-Pinedo, F., et al. (2021). An HPC multi-physics framework for next-generation simulations. Scipedia, 17(2), 45–56.

Snir, M., et al. (2020). Checkpointing and fault-tolerance for exascale systems. ACM Computing Surveys, 53(3), 1–38.

Trebotich, D., et al. (2023). Exascale-coupled multiphysics frameworks: Coupling flow and reactive transport simulation. Frontiers in High-Performance Computing, 2, 115–127.

Wall, W. A., & Gee, M. W. (2019). Fluid–structure interaction at the limits of computational power. Computational Mechanics, 64(3), 777–793.