



Rescale at NAFEMS World Congress 2025

Explore Rescale's accepted papers from NAFEMS World Congress 2025, showcasing innovations in cloud, AI, and HPC.

Table of Contents

- 3 Accelerating Scientific Workflows with Domain-Specific Hardware: GPUs, Arm Chips and Beyond
- 11 Challenges and Opportunities in Cloud-Based Simulation - An Engineer's Perspective
- 25 Safeguarding Engineering IP in the Cloud: Strategies for Secure Global Collaboration
- 35 Orchestrating Hybrid HPC Environments: Strategies for Data Gravity and AI-Ready Datasets
- 44 Digital Thread Foundations for Accelerated Multi-Disciplinary CAE Workflows
- 53 Reshaping Simulation Data for an AI Future
- 61 Leveraging LLMs for Automated Post-Processing of Simulation Output Logs

Accelerating Scientific Workflows with Domain-Specific Hardware: GPUs, Arm Chips and Beyond

Dr Sam Zakrzewski PhD, Rescale (Denmark)

Romain Klein, Rescale (France)

Abstract

The demand for high-performance computing (HPC) is expanding as scientific and engineering challenges grow increasingly complex. Traditional CPU-based architectures, while versatile, often struggle to efficiently handle specialised workloads such as machine learning, computational fluid dynamics (CFD), and molecular dynamics. To address this, the integration of domain-specific hardware accelerators like NVIDIA GPUs and Arm chips has emerged as a game-changer, enabling unparalleled performance and efficiency for targeted applications.

This paper delves into the role of domain-specific hardware accelerators in revolutionising scientific workflows. We focus on how specialised architectures, available on Rescale's intelligent cloud HPC platform, empower researchers and engineers to leverage cutting-edge hardware tailored to their workloads. By combining GPUs for compute-intensive tasks with aarch64 architectures for energy-efficient operations, users can achieve optimal performance while addressing cost and sustainability goals.

A critical aspect of this discussion involves the optimisation of workflows for hybrid and heterogeneous computing environments. Integrating domain-specific accelerators requires not only hardware availability but also seamless software orchestration to manage data flows, scheduling, and execution. Rescale's platform addresses these challenges by offering a unified environment where users can dynamically select the most suitable hardware configurations based on workload requirements. This flexibility is particularly impactful for industries ranging from aerospace to pharmaceuticals, where precision and efficiency are paramount.

The paper also explores practical use cases, such as leveraging GPUs for AI-driven simulation post-processing or Arm-based chips for low-power, high-throughput scenarios. These examples illustrate how hardware accelerators can drastically reduce time-to-solution and cost, enabling organisations to push the boundaries of innovation. Additionally, we will discuss the importance of workload profiling and benchmarking to ensure optimal hardware utilisation, drawing on real-world insights from Rescale's platform.

Attendees will gain an understanding of the technical and operational considerations involved in adopting domain-specific hardware, from software compatibility to deployment strategies in cloud-based environments. We will highlight how Rescale simplifies these complexities, allowing users to focus on their core scientific and

engineering objectives.

Through this exploration of domain-specific accelerators, we aim to demonstrate how specialised architectures are not just advancing performance but also reshaping how organisations approach HPC, paving the way for new breakthroughs in research and industry.

1. Introduction

High-performance computing is the backbone of modern engineering and scientific research. As the complexity of simulations grows, so does the demand for efficient and scalable compute resources. Traditional CPU-based systems, governed by Moore's Law, have struggled to meet this demand. Specialized architectures, such as GPUs and domain-specific accelerators, offer a solution, delivering exponential performance improvements [1]. Cloud-based platforms have emerged as a pivotal enabler, providing engineers with seamless access to these technologies while optimizing costs and sustainability.

Over the past decades, engineering simulations have undergone a profound transformation. From initial wireframe models to sophisticated, multi-physics simulations, the computational demands have increased manifold. The limitations of traditional HPC infrastructures have spurred the need for innovative solutions. Today, engineers require platforms that not only deliver raw computational power but also adapt to specific workloads and reduce environmental impacts.

Rescale addresses these needs by offering a unified platform that integrates state-of-the-art hardware and intelligent software. By leveraging cloud-based infrastructure, Rescale eliminates the bottlenecks associated with static, on-premises systems. Engineers can dynamically scale resources to match the complexity and urgency of their projects, ensuring that insights are derived faster and more efficiently. This capability is particularly critical in industries such as aerospace, automotive, and life sciences, where time-to-market and precision are paramount.

The integration of AI and machine learning has redefined how simulations are performed. AI-assisted workflows not only enhance accuracy but also automate repetitive tasks, freeing engineers to focus on innovation. Rescale's platform supports these advancements, offering pre-configured environments for deploying AI-driven simulations. This symbiosis of advanced hardware and intelligent algorithms marks a new era in digital engineering.

2. Trends in the Industry

Sustainability and Energy Efficiency

Emerging architectures promise up to five times better energy and performance efficiencies compared to traditional on-premises solutions [2]. This shift aligns with global sustainability goals, encouraging organizations

to adopt cloud-based HPC resources that are not only faster but also greener. Companies now prioritize reducing their carbon footprints while optimizing computational workloads, and cloud-based platforms play a crucial role in achieving these objectives.

Collaborative Research and Data Sharing

Approximately 70% of organizations globally engage in collaborative R&D, often bridging academia, suppliers, and service providers. This collaboration ensures a steady exchange of knowledge and innovation across sectors. Platforms like Rescale facilitate seamless data sharing and collaborative workflows, such as multiple users accessing simulation output stored in secure, centralized location, enabling geographically dispersed teams to work together efficiently. The rise of interdisciplinary projects further underscores the importance of robust data-sharing mechanisms, which enhance the development of next-generation products and solutions.

Accelerated Computing and AI-Driven R&D

The performance of GPUs, RISC architectures, and similar accelerators is advancing at a rate double that of traditional x86 CPUs. These technologies, coupled with AI and machine learning, have transformed research and development processes across industries. AI-driven R&D enables faster prototyping, predictive modeling, and optimization, cutting development cycles by up to 50%. This progress not only accelerates time-to-market but also enhances product quality and operational efficiency, ensuring companies stay ahead in competitive markets.

3. Specialized Architectures in Engineering Simulations

Advantages of Domain-Specific Accelerators

Domain-specific hardware accelerators, including GPUs, FPGAs, TPUs, and quantum computing platforms, offer tailored solutions for various simulation needs. For instance, GPU-based solutions excel in parallelized workloads such as computational fluid dynamics (CFD), while high-memory-bandwidth architectures are ideal for finite element analysis (FEA). These accelerators not only enhance computational speeds but also reduce operational costs by optimizing resource allocation. By matching hardware capabilities to specific tasks, organizations can achieve significant gains in productivity and performance [1,2,3].

Case Study: ANSYS Fluent

Benchmarks show significant performance gains using Nvidia GPUs, enabling faster external flow simulations with millions of cells. This improvement has allowed teams to conduct more iterations in less time, enhancing design accuracy.

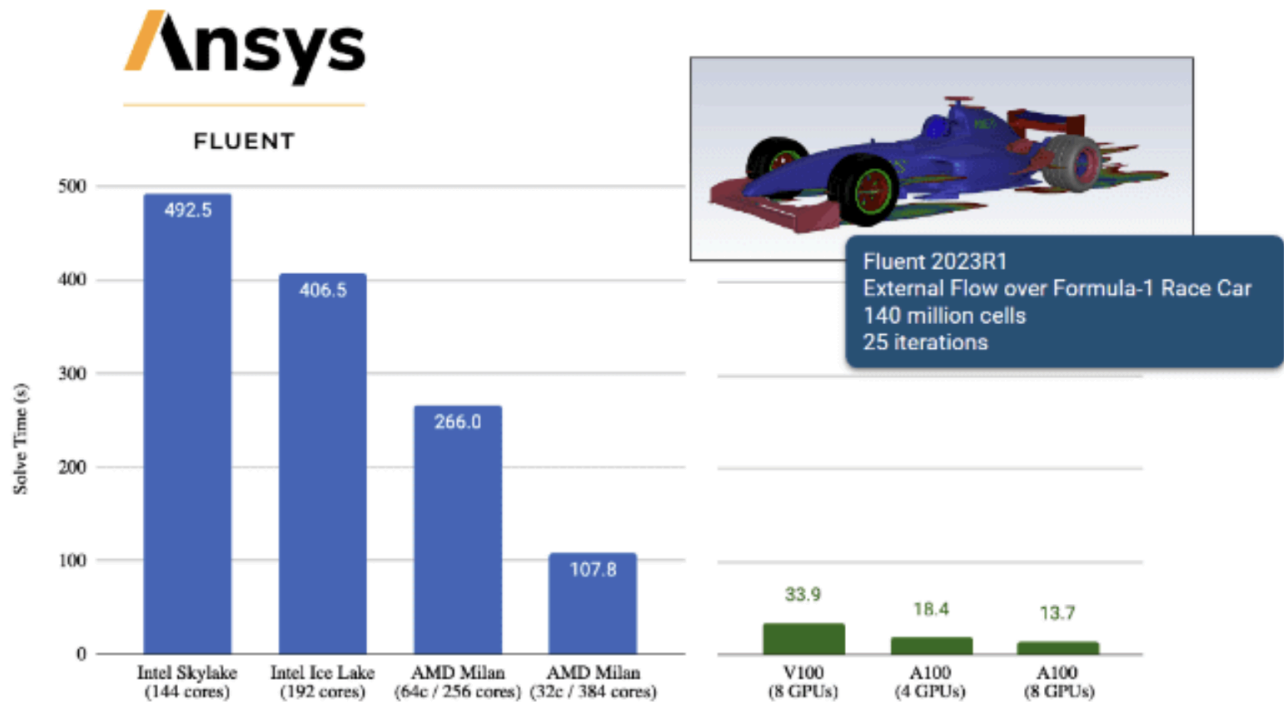


Figure 1 Comparison of CPU and GPUs performance on ANSYS Fluent [5]

The workflow is as follows:

- **Data Preparation**
Preprocesses the geometry and mesh data, then upload it to Rescale's cloud storage.
- **Hardware Choice**
Select a Rescale coretype equipped with Nvidia A100 GPUs, known for their high performance in parallel computing tasks.
- **Simulation Execution**
The CFD solver code utilizes CUDA for parallel processing. The ANSYS Fluent solver is configured to take advantage of the GPU's capabilities.
- **Optimization and Scaling**
Using Rescale's Performance Profile tools, identify performance bottlenecks and optimize the workflow. Additional options to scale the simulation across multiple GPUs to further reduce computation time for a cost-efficient solution

Case Study: OpenFOAM

Rescale supports aarch64 and x86 architectures, offering flexibility in deployment and optimized performance for specific workloads [3]. OpenFOAM users benefit from Rescale’s ability to provide tailored computational environments, streamlining their simulation processes and boosting overall efficiency.

Performance Scale Index

Higher is better

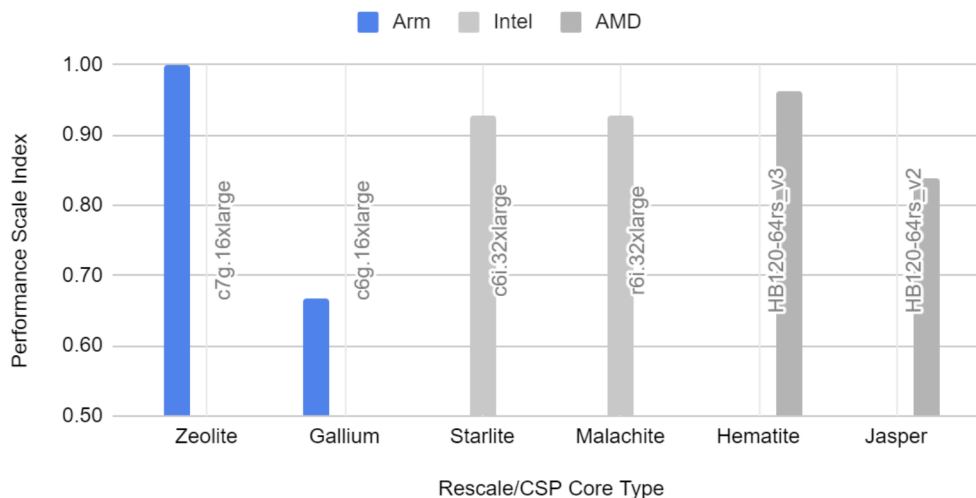


Figure 2 Comparison of aarch64 and x64 architectures on OpenFoam [6]

Here the data shows a favourable comparison between Arm architecture (Zeolite AWS Graviton 3) and x86 based cloud systems

Benchmarks

Hardware benchmarks highlight the performance of the latest accelerators, such as Nvidia’s A100 GPUs and Arm’s Graviton processors. These benchmarks validate the cloud’s ability to deliver best-in-class performance and cost efficiency. For example, simulations conducted on the A100 GPUs demonstrated up to a 4x [4] improvement in processing times, underscoring the advantages of using specialized architectures for intensive engineering tasks.

4. AI-Driven Engineering

Impact of AI/ML

AI/ML has emerged as a transformative force in engineering. From training neural networks on physics data to deploying pre-trained models for predictive simulations, AI accelerates insights and optimizes workflows. These capabilities allow engineers to identify potential issues early in the design process, reducing costly revisions and ensuring products meet stringent quality standards. By integrating AI into simulation workflows, Rescale empowers organizations to transition from reactive to proactive engineering methodologies.

AI Physics Breakthroughs

AI enables computational reducibility, unlocking accurate predictive physics that traditional methods cannot achieve. Pre-trained physics models and accelerated computing architectures further enhance these capabilities, reducing runtime and computational costs. Recent breakthroughs include the development of neural networks capable of solving complex fluid dynamics problems in real time, offering engineers unprecedented speed and precision in their analyses.

Design Optimization

AI integrates seamlessly into simulation workflows, automating tasks such as drag coefficient calculations and design validation. Engineers can rapidly evaluate multiple designs, shortening the design cycle and improving product quality. Additionally, AI-driven optimization tools can explore vast design spaces, identifying configurations that maximize performance while minimizing material usage and costs. This holistic approach to design ensures that products are not only high-performing but also sustainable.

5. Digital Transformation and Ecosystem Integration

Elastic Cloud Workstations

Rescale's Elastic Cloud Workstations offer a powerful, web-based platform for simulation workflows. Engineers can access full application catalogs, deploy secure virtualization solutions, and collaborate with colleagues in real time. These workstations provide the flexibility to pre-process, solve, and post-process simulations from any location, ensuring productivity remains uninterrupted. The integration of cutting-edge hardware and intuitive software enables seamless transitions between different phases of simulation, enhancing overall efficiency.

End-to-End Workflows

The platform completes the digital thread by integrating with tools for product lifecycle management (PLM), workflow orchestration, and schedulers. This integration ensures a cohesive ecosystem for engineering simulations, from data preprocessing to post-processing. Rescale's support for diverse applications and hardware configurations allows teams to customize their workflows, aligning them with specific project

requirements [5]. This adaptability is critical in dynamic industries where innovation and agility are key to success.

6. Rescale's Role in HPC

Platform Capabilities

Rescale offers a cloud-based HPC platform that integrates seamlessly with the latest hardware technologies, enabling organizations to harness diverse compute resources on-demand [7]. Engineers can pre-process, solve, and post-process simulations using a web-based GUI, fostering accessibility and collaboration. This approach eliminates the logistical and financial constraints of maintaining on-premises HPC infrastructure, providing engineers with flexibility and scalability. The platform's intuitive interface and performance insights allow users to make data-driven decisions, optimizing resources for maximum efficiency.

Strategic Partnerships

Rescale's collaborations with industry leaders like Nvidia [10] and Arm [9] bring cutting-edge hardware to its users. Nvidia's contributions, including integration with the NGC Catalog and DGX Cloud, enable efficient AI and simulation workloads. Similarly, Arm-powered instances optimize chip design and verification processes, showcasing the versatility of domain-specific hardware. These partnerships ensure that Rescale's platform remains at the forefront of technological advancements, offering users unparalleled performance and cost-efficiency.

7. Conclusion and Future Directions

Cloud-based platforms represent a paradigm shift in HPC for engineering simulations. By enabling access to domain-specific accelerators and AI-driven workflows, they empower organizations to achieve faster, more sustainable, and cost-effective results. As the adoption of specialized architectures continues to grow, these versatile powerhouses will remain at the forefront, driving innovation and shaping the future of engineering and scientific research. Future developments may include deeper integration with quantum computing platforms, expanded AI capabilities, and enhanced support for interdisciplinary collaborations, ensuring that Rescale continues to meet the evolving needs of its users.

8. References

- [1] Nvidia. "Heeding Huang's Law: Video Shows How Engineers Keep the Speedups Coming." <https://blogs.nvidia.com/blog/huangs-law-daily-hot-chips/>
- [2] ANSYS. "Volvo Cars Leverages Ansys and NVIDIA GPUs to Accelerate CFD Simulations by 2.5x for the EX90 Electric Vehicle" <https://www.ansys.com/news-center/press-releases/3-18-25-ansys-volvo-accelerate-cfd-simulation-with-nvidia-gpu>
- [3] Arm. "Ansys Fluent® and Ansys LS-DYNA® demonstrates leading performance on Arm" <https://community.arm.com/arm-community-blogs/b/servers-and-cloud-computing-blog/posts/ansys-fluent-ansys-ls-dyna-on-arm>
- [4] ANSYS. "Unleashing the Full Power of GPUs for Ansys Fluent, Part 2" <https://www.ansys.com/blog/unleashing-the-full-power-of-gpus-for-ansys-fluent-part-2>
- [5] Rescale. "Leveraging Specialized Architectures with Domain-Specific Hardware Accelerators: Nvidia GPUs and Arm Chips on Rescale." <https://rescale.com/blog/leveraging-specialized-architectures-with-domain-specific-hardware-accelerators-nvidia-gpus-and-arm-chips-on-rescale/>
- [6] FOAM Iberia 2023. "Leveraging Arm Architecture And Rescale Cloud Hpc Platform For Enhanced Openfoam Performance: A Comparative Analysis" <https://2023.foam-iberia.eu/program/> Session T3
- [7] Rescale. "Cloud Computing Infrastructure for HPC." <https://rescale.com/platform/hpc-ai-architectures/>
- [8] Rescale. "GPU Center of Excellence." <https://rescale.com/solutions/by-use-case/gpu-coe/>
- [9] Arm. "Arm HPC Ecosystem." <https://www.arm.com/solutions/high-performance-computing>
- [10] Nvidia. "NVIDIA GPU Solutions for HPC." <https://www.nvidia.com/en-us/data-center/gpu-acceleratedapplications/high-performance-computing/>

Challenges and Opportunities in Cloud-Based Simulation - An Engineer's Perspective

John William, Rescale (Netherlands)

Romain Klein, Rescale (France)

Abstract

The transition from on-premise to cloud-based high-performance computing (HPC) is a growing trend in engineering simulation, driven by advancements in cloud technology. This shift offers scalability, flexibility, and access to cutting-edge hardware without significant upfront investment. However, it also presents unique challenges that require strategic planning.

This paper, from a seasoned simulation engineer with over 25 years of experience, provides practical advice and technical strategies for migrating simulation workloads to the cloud. Key topics include hardware configuration, data management, workflow optimization, cost management, and license allocation. The discussion also covers collaborative dynamics and data-sharing in cloud environments, with real-world examples illustrating how to overcome common obstacles and maximize the benefits of cloud-based simulation.

By leveraging cloud technologies, engineers can enhance efficiency, reduce costs, and accelerate innovation. This presentation aims to equip engineers with the knowledge and tools to navigate the complexities of cloud-based simulation effectively.

1. Introduction

The landscape of engineering simulations is undergoing a transformative shift as organizations migrate from on-premise HPC to cloud-based solutions. This transition offers scalability, flexibility, and cost-efficiency but presents unique challenges that require strategic planning. This paper provides actionable insights for engineers navigating this change.

High-Performance Computing (HPC) has traditionally relied on on-premise systems, which required significant capital investment and maintenance. Over the years, advancements in cloud computing have made it a viable alternative, offering similar or even superior performance with added benefits of scalability and flexibility[1]. The evolution from on-premise to cloud-based HPC has been driven by the need for more agile and cost-effective solutions.

The adoption of cloud-based HPC is accelerating. According to recent studies, the cloud HPC market is expected to grow at a compound annual growth rate (CAGR) of 18.1% from 2022 to 2027, reaching \$14.1 billion by 2027[2]. This growth is fueled by the increasing demand for scalable computing resources, the rise of AI and machine learning workloads, and the need for faster innovation cycles. Companies are leveraging cloud HPC to handle peak workloads, reduce time-to-market, and enhance collaboration across geographically dispersed teams.

Looking ahead, cloud-based HPC is poised to become even more integral to engineering simulations. The future will likely see greater integration of AI and machine learning, enabling more sophisticated simulations and predictive analytics. Additionally, advancements in cloud technologies, such as quantum computing and edge computing, will further expand the capabilities and applications of cloud HPC. As organizations continue to embrace digital transformation, cloud HPC will play a crucial role in driving innovation and maintaining competitive advantage.

The shift to cloud-based HPC represents a significant change for engineers accustomed to on-premise systems and schedulers. Many engineers, with backgrounds in science or engineering, are new to cloud technology and face a steep learning curve.

In on-premise environments, costs are typically capital expenditures (CapEx) and not directly visible to engineers. However, cloud computing operates on an operational expenditure (OpEx) model, where engineers must manage and optimize budgets in real-time. This shift requires engineers to develop new skills in cost management and budgeting.

Moving large datasets to and from the cloud can be challenging due to data gravity and transfer costs. Data gravity refers to the tendency of data to attract additional data and applications, making it difficult and costly to move. Engineers must consider these factors when planning data migration and storage strategies.

In on-premise systems, shared file systems facilitate easy sharing of simulation results among colleagues. In the cloud, engineers must navigate different tools and platforms for collaboration, which can introduce complexity. Cloud-based solutions should offer capabilities for sharing and collaborating on simulation models, but most of the time they require adaptation to new workflows.

On-premise environments often have well-established license queuing and management systems. In the cloud, managing software licenses can be more complex, with different models such as floating licenses and subscription-based licenses. Engineers need to adapt to these new licensing models and ensure efficient resource utilization.

Automating simulations in the cloud involves integrating various tools and platforms, which can be challenging for engineers used to on-premise systems. Cloud-native approaches to simulation data automation can help streamline these processes, but they require a shift in mindset and workflows.

This paper looks into some of these challenges and provides insights into how a simulation engineer adapts these in the context of running simulations on the cloud HPC systems.

2. Hardware Configuration Challenges

Cloud environments offer diverse hardware options, such as CPU and GPU instances. Engineers must evaluate

simulation needs to select optimal configurations. For instance, simulations requiring heavy parallel processing may benefit from GPU instances, while others might favor CPU efficiency.

The rapid advancement and availability of specialized HPC hardware in the cloud have created both opportunities and challenges for R&D and engineering teams. These teams, tasked with developing advanced products, now have access to a wide array of cloud-based HPC resources to power their computation-intensive applications. However, the explosion of specialized hardware options, coupled with the increasing complexity of applications, necessitates a strategic approach to hardware configuration.

Benchmarking is essential for optimizing HPC operations in the cloud. By systematically evaluating commonly used computer-aided engineering (CAE) applications, organizations can identify the most suitable hardware and software configurations for their specific business goals. This data-driven approach helps capture more value from the latest cloud-HPC technologies, ensuring that resources are used efficiently and effectively.

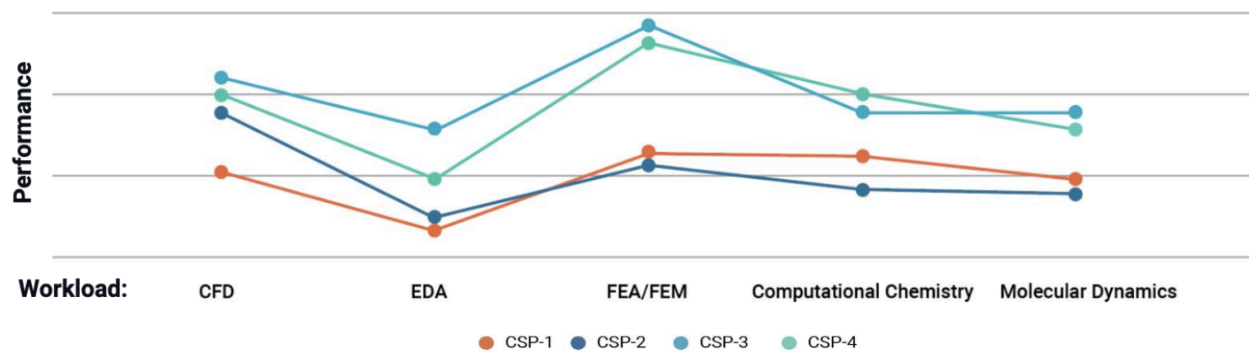


Figure 1: Average Cloud Provider Performance Across Applications

Simulation engineers and HPC leaders must consider whether they are using the best cloud architectures for their specific workloads, optimized for cost, speed, or scale. They need strategies to keep up with the latest technologies across multiple providers and leverage metrics and data sources to maximize the performance and efficiency of their HPC resources. Businesses can increase workload performance by 30% and reduce overall simulation costs by 20% by selecting more optimal hardware available in the cloud [3].

The variety of specialized HPC hardware options in the cloud has grown significantly, including x86, GPU, ARM, RISC-5, and FPGA chip types [3]. This diversity allows engineers and scientists to select the most appropriate hardware for each specific workload. However, this abundance of choices can be overwhelming, making it challenging to make informed decisions without the right tools and intelligence.

The complexity and variety of applications used in HPC have also increased. Organizations often use multiple digital R&D and engineering applications, each with specific hardware requirements. Continuous benchmarking and optimization of workloads are crucial to ensure that the most suitable hardware is used, leading to significant performance improvements and cost savings.

High-impact HPC decisions require access to up-to-date benchmarking data. By analyzing performance across various cloud hardware options, organizations can improve workload efficiency, time to solution, and overall

return on investment. Continuous benchmarking helps organizations stay ahead of the curve, adapting to new technologies and optimizing their HPC operations.

Benchmarking over time reveals trends in cost-performance improvements. Organizations can achieve significant gains by switching to the best available cloud hardware, realizing up to a 30% year-on-year increase in cost-performance. This continuous improvement underscores the importance of having multi-cloud intelligence and flexibility to capitalize on new hardware releases.

In summary, the hardware configuration challenges in cloud-based HPC require a strategic approach to benchmarking and optimization. By leveraging performance intelligence and continuously updating their hardware and software stack, organizations can maximize the efficiency and effectiveness of their HPC resources, driving innovation and maintaining a competitive edge.

3. Data Management in Cloud Environments

As a simulation engineer, I've seen firsthand how computational science and engineering have revolutionized our field, but they've also brought significant data management challenges. Our product development teams are now flooded with data from various sources, including product design, simulation, modeling, and machine learning applications. This influx of data results in uniquely complex data and file types that hold essential information about the assumptions, parameters, and results of our simulations.

Without proper data management practices, we struggle to track and share both the content and context of this simulation data. It's not uncommon for us to accumulate several terabytes or even petabytes of historical computational science and engineering data. This data often becomes difficult to access and share because its contents are not well-documented, yet it cannot be deleted for the same reason. This situation makes it extremely challenging to ensure the reproducibility of simulations, which is crucial for understanding past work and its implications on design decisions.

To address these challenges, we need to gain mastery over this growing array of complex yet essential data. This requires a shift from ad hoc and fractured data management practices to a unified and strategic approach. Understanding the complexity of data types and sources in simulation, modeling, and AI/ML is the first step toward achieving this goal. We utilize and generate a wide variety of data types, including geometries, materials, CAE models, software automation scripts, databases, and results such as KPIs, plots, graphs, and visualizations. These data types are stored across various services, including object storage, fast parallel file systems, local disks, cloud file systems, databases, and data lakes.

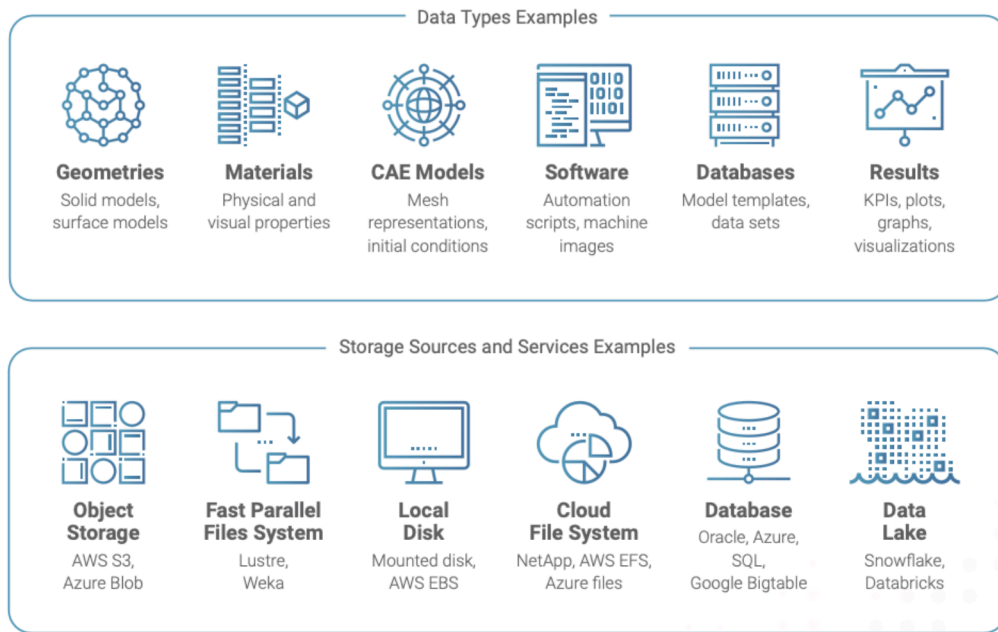


Figure 2: Complexity of Data Types and Sources in Simulation, Modeling, and AI/ML

While digital approaches to engineering are becoming mainstream, the digital transformation of R&D programs is still held back by data management challenges. Establishing consistent digital practices across teams and processes involved in new product development is critical, yet difficult to achieve.

One major challenge we face is data fragmentation. Information is often stored in different systems, making it difficult to access. For example, a shared drive used by one team for a particular simulation software may not be accessible to others. This lack of a coherent storage strategy leads to data being scattered across various file systems and storage environments, further exacerbated by the adoption of various cloud services.

Another significant challenge is data disconnection. Accessed data often lacks context, making it less useful. For instance, a random simulation output file without a history of analysis provides little value. Capturing the intention and history of simulations is crucial for effective data and knowledge management. However, we typically focus on our R&D tasks rather than on metadata capture and enforcement, leading to disconnected data.

These challenges result in several issues. A lack of data sharing leads to duplication of efforts and isolated insights, creating islands of analysis. Without a holistic view of overall progress, insights on product performance and dependencies are difficult to obtain, leading to incomplete visibility. Additionally, without clear records of how product design decisions were made, assessing and improving product development processes becomes challenging, resulting in opaque product decisions.

Addressing these challenges requires a modern approach to data and analytics governance. We must develop strategies to manage data fragmentation and disconnection, ensuring that data is accessible, contextualized, and useful for decision-making.

As simulation, modeling, AI, and other digitization initiatives generate a growing volume of complex data, we must shift to a new approach in how we manage our R&D computing data. Digital engineering and product

development teams need to create a cohesive data-driven approach to their workflows and processes. This starts with foundational practices that bring teams and insights together to produce better products and operate more efficiently. Information must be accessible to authorized users regardless of where or how it is stored, and product data should be accompanied by information on how and why the data was created, what elements were in scope, and how the data was used.

As simulation, modeling, and AI/ML activities increase, it is important that our storage architecture can scale cost-effectively with data and user growth. Establishing a data architecture that is multi-tiered allows each engineering team to utilize a variety of data management services to meet specific digital R&D needs and business objectives. Federated storage enables multiple teams to easily access and share commonly used files across an array of independent storage platforms. Governance ensures that policy-based storage automations for regional data locality, industry compliance, or cost-performance requirements are in place. Durability ensures that critical data is persisted and regularly backed up in multiple regions to support quick data recovery and mitigate the risk of data loss.

Shared context between teams means that anyone can easily access details about engineering data they have access to. Metadata management for simulation, modeling, and AI/ML analysis can be accomplished in different ways based on organizational preferences, such as free-form commenting, flexible tagging, controlled taxonomies, and prescriptive forms. This additional information ensures that the intention of each digital activity is captured consistently and the interpretation of data is accurately depicted for clear handoffs to other teams.

To improve the accuracy and agility of our R&D efforts, we should aim to integrate capabilities that ensure metadata is added seamlessly within end users' existing tools and workflows. Providing flexibility in data capture and searchability, enforcing specific metadata requirements to establish process consistency and ensure data quality, enabling teams to search, filter, and perform analytics on metadata from related studies, and allowing import and export of metadata using APIs for third-party tools are all critical steps.

By effectively managing our engineering and simulation data, we can digitally transform our R&D processes. This transformation can improve R&D velocity, efficiency, and product quality, empowering us with the information we need when we need it. In this new world of connected and contextual collaboration, all teams have access to the shared insights they need to innovate, regardless of where or how the data is stored.

4. Cost Optimization Strategies

Balancing performance with budget is key in cloud-based HPC environments. Engineers must employ strategies such as using reserved instances for cost savings, spot instances for flexible use, and on-demand instances for immediate needs. Effective cost optimization involves leveraging advanced tools and best practices to ensure that computational resources are used efficiently without compromising performance.

One of the most effective ways to manage and optimize costs is through the use of cost prediction tools. These tools enable organizations to benchmark and assess the performance of various computing architectures, helping them identify the most cost-effective configurations for their specific workloads. Automated benchmarking capabilities allow for the evaluation of performance, cost, energy consumption, and carbon footprint across different infrastructure options.

For instance, performance profiling tools can automate the benchmarking of workloads and datasets, providing

insights into the optimal hardware and software configurations. This automation helps organizations capture significant cost savings and performance improvements for computational engineering, machine learning, and generative AI workloads. By identifying the best configurations, businesses can save thousands or even millions of dollars while reducing run times and accelerating innovation.

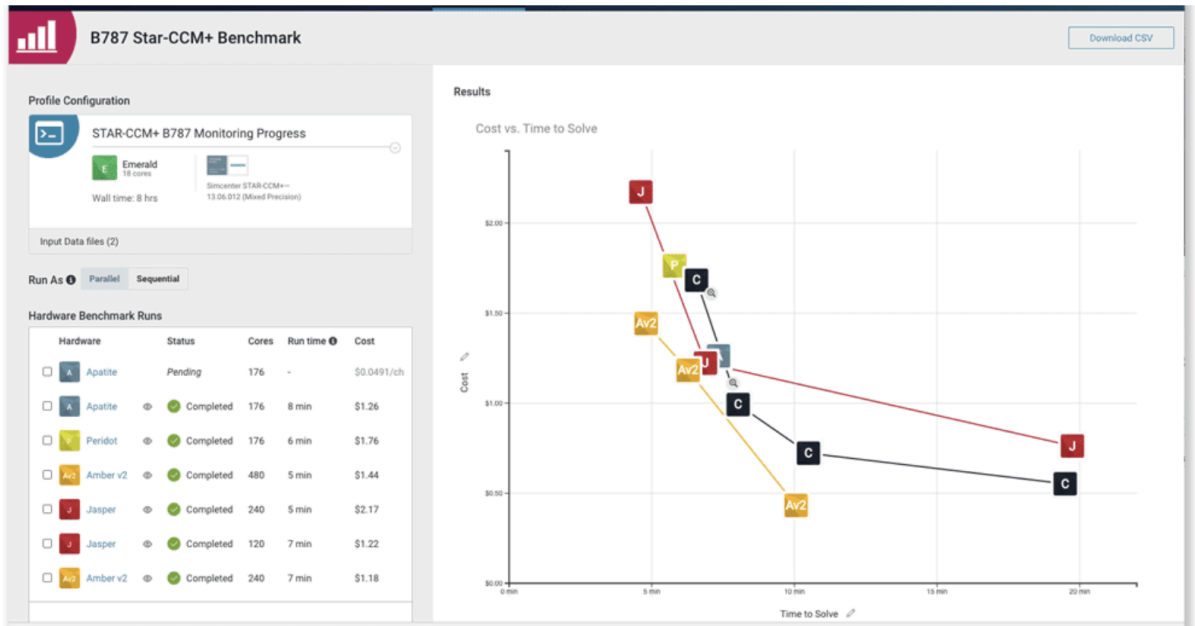


Figure 3: Examples of an intelligent performance profiling tool

Manual benchmarking has traditionally been a common practice, but the rapid growth of specialized microprocessor chips and HPC infrastructure services from cloud providers has made it increasingly difficult to identify the best configurations manually. Automated tools simplify this process, allowing organizations to quickly and easily benchmark applications across multiple types of microprocessors and core configurations using their own simulations and models.

These tools work in conjunction with AI-driven recommendation engines, which suggest the best computing architectures for benchmarking based on the specific application. Engineers can then comprehensively assess the best architecture options through an intuitive console, even if they are not highly technical users. Promising configurations can be saved to job templates, ensuring that the best infrastructure configuration is deployed every time.

With near real-time awareness of the costs and performance of any cloud computing architecture, performance management becomes a seamless and continuous process. This approach eliminates the need for time-consuming, manual benchmarking efforts and provides organizations with the information they need to manage costs effectively while driving innovation.

Selecting the right type of instance based on workload type is another critical aspect of cost optimization. Reserved instances offer significant cost savings for predictable workloads by providing a lower hourly rate in exchange for a commitment to use the instance for a specified term. Spot instances, on the other hand, allow for flexible use and can be significantly cheaper than on-demand instances, making them ideal for non-critical workloads that can tolerate interruptions. On-demand instances provide the flexibility to scale resources up or down based on immediate needs, ensuring that engineers have access to the necessary computational power without long-term commitments.

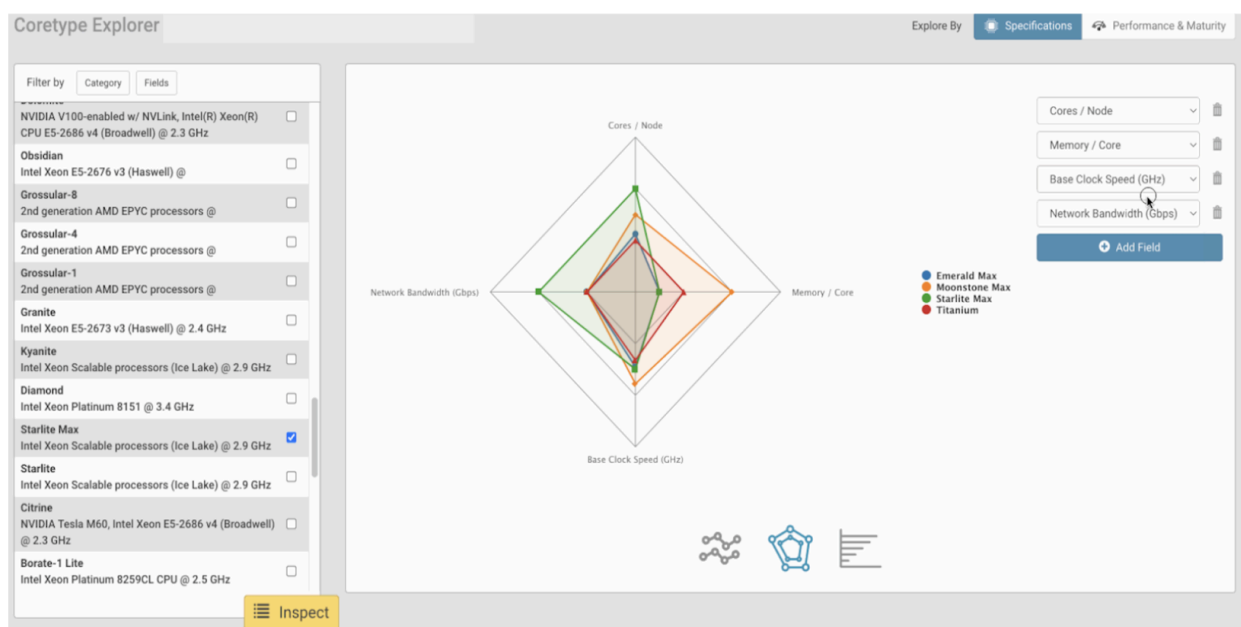


Figure 4: Examples of an instance selection tool

By leveraging these strategies and tools, organizations can balance performance with budget effectively, ensuring that their HPC resources are used efficiently and cost-effectively. This approach not only reduces overall simulation costs but also enhances the ability to innovate and respond to changing computational demands.

While using spot instances offers a cost-effective way to run simulations, careful consideration should be given to their usage for long-running and time-critical simulations. Spot instances are provided by cloud service providers at attractive prices, but these instances can be reclaimed by the provider with limited warning if there is a rising demand with better cost offerings. In such cases, simulations are terminated immediately, potentially resulting in the loss of simulation data and time. In addition to hardware costs, one must also consider software licensing costs and the valuable time engineers spend running the simulations. Therefore, it is important to have preventive mechanisms in place to mitigate the loss of efficiency due to spot terminations.

One common practice in engineering simulations that can address this issue is checkpointing. Traditionally, checkpointing is used in on-premise engineering simulations to prevent data loss in case of running out of disk

space or other hardware failures. It is also used to run divergent load cases from a common starting point in simulation history or to transfer the results of one simulation to another type of simulation.

In the context of cloud HPC, checkpoint restarting schemes are particularly useful when using spot instances. Existing checkpointing schemes can be augmented to create reusable snapshots of results at predetermined intervals. Cloud HPC orchestration schemes can be set up so that engineering simulations can continue on a newly created cluster from the last available checkpoint result from the previously failed spot instance.

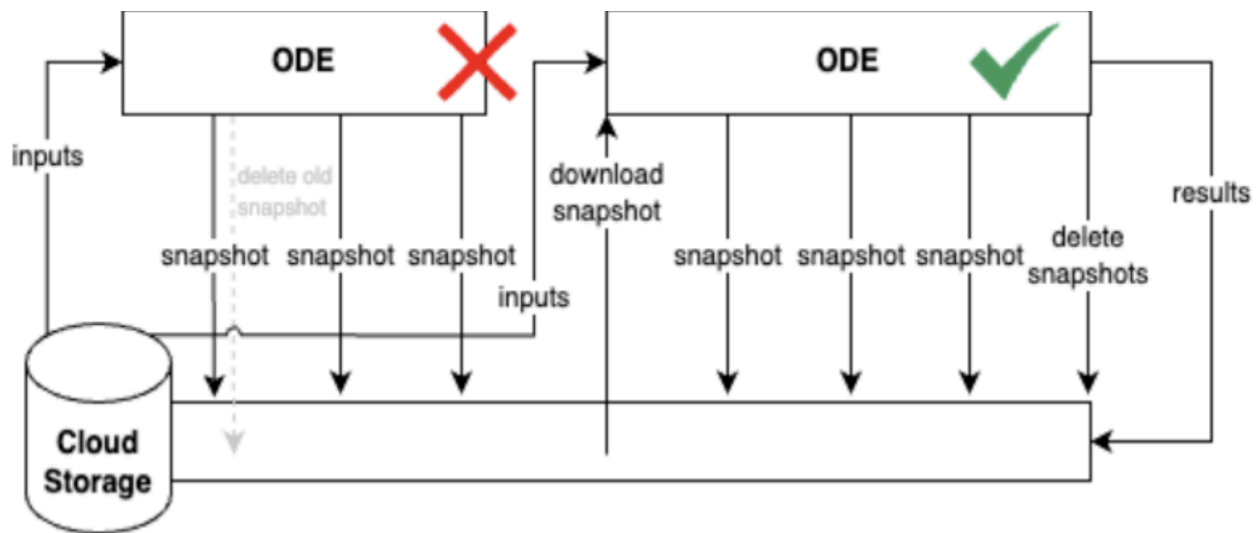


Figure 5: Examples of checkpoint restarting scheme in a Cloud HPC simulation.

The Cloud HPC platform should be flexible enough to incorporate orchestrating the restart behavior and managing other aspects of the simulation, such as licensing and data storage, to ensure the continuity of the simulation.

5. License Management

Cloud environments pose challenges with software licensing. Approaches include floating licenses and subscription models to efficiently manage resources across distributed teams.

License management is a critical component of cloud-based HPC systems, ensuring that software licenses are efficiently utilized and shared across both on-premise and cloud environments. Effective license management allows organizations to maximize their investment in software licenses while maintaining compliance and optimizing resource usage.

The cloud HPC system must complement and integrate seamlessly with existing on-premise hardware by sharing a common set of licenses. It should facilitate license usage across different cloud service providers, enabling a secure connection to on-premise hosted license servers or providing the option to host licenses directly on the cloud HPC system. This flexibility is essential to accommodate different types of license servers

required by various software applications.

Different connection methods, such as IP address/port forwarding, establishing an SSH connection, or setting up a VPN connection between the license servers and the cloud HPC system, should be considered. These methods ensure secure and reliable access to licenses, regardless of the deployment environment. Additionally, the license calculation schemes should be automatically adjusted based on the types of cloud hardware chosen by the users, ensuring accurate and fair allocation of resources.

Another important aspect of license management is the allocation of licenses per user, project, or group. This allocation should be integrated into the cost and budgeting of cloud HPC resources, allowing organizations to track and manage expenses effectively. Suitable license queuing schemes should be in place to manage the order in which simulation jobs are executed and to ensure the timely release of licenses once simulations are completed.

Furthermore, the licensing system should provide enough flexibility to allocate licenses based on geographic regions and time zones. This capability ensures that licenses are well-utilized across global user teams working in different locations around the world. By implementing these strategies, organizations can optimize their license usage, reduce costs, and enhance the overall efficiency of their cloud-based HPC systems.

6. Collaborative Dynamics

Cloud platforms enable seamless teamwork through shared storage, version control systems, and real-time collaboration tools, fostering efficiency among geographically dispersed teams.

Effective collaboration is essential in modern engineering simulations, especially as teams become more distributed and projects grow in complexity. Cloud-based HPC systems facilitate connected simulation collaboration and data sharing, enabling distributed teams and value-chain partners to work together seamlessly on product R&D. By unifying silos of analysis across the organization with shared access and context on computing and data resources, data quality and process efficiency in complex multidisciplinary projects can be significantly improved.

Collaborating directly on simulation models with shared resources such as compute jobs, interactive workstations, and files eliminates the need for ad-hoc sharing of resources. Streamlined, policy-based access for the entire team ensures that everyone has the necessary tools and data at their fingertips, enhancing productivity and reducing the risk of errors. Shared computing workspaces provide a centralized environment where team members can collaborate effectively, regardless of their physical location.

Capturing simulation insights and building shared context is crucial for accelerated handoffs and informed product decisions. Administering workspaces with templated custom fields for governed simulation data allows for the filtering and searching of tagged data, improving the discoverability of simulation data. By sharing context, intent, and outcomes, teams can ensure that all relevant information is available for decision-making, reducing the time spent on data management and preventing duplicate work.

Real-time collaboration directly on product models is facilitated through comments and notifications. Team members can seek feedback, add additional context, and record decisions, ensuring that important information is captured and shared. Features such as pinning important comments, mentioning team members, and linking to other simulation jobs enable seamless knowledge sharing and enhance the overall efficiency of the

collaboration process.

Accelerating repetitive workflows and best practices with templates for jobs and workstations increases R&D productivity. Policy-driven automations across workspaces and projects streamline the onboarding of new users and improve overall process efficiency. By standardizing common tasks and configurations, teams can focus on innovation rather than administrative tasks.

Connecting disparate data sources to provide uniform data access to collaborating teams and value chain partners is essential for effective collaboration. From simulation input/output files to open-source data sets for model training, engineers and scientists can synchronize and organize project files seamlessly. A unified storage fabric ensures that all team members have access to the data they need, when they need it, regardless of the underlying storage technology.

Ensuring traceable data and decisions is critical for maintaining the integrity of the simulation process. Real-time documentation of decisions with comments and notifications, along with linking simulation jobs with in-line references to other simulation jobs, provides a clear audit trail. This traceability allows teams to search and find simulation context, inputs, and findings when needed, ensuring that all decisions are well-documented and easily accessible.

Improving the quality of simulation insights through policy-driven governance, analytics, and simulation templates ensures data consistency and reliability. Programmatically synchronizing with sources of truth, such as PLM and SPDM systems, helps maintain accurate and up-to-date information. This governance framework supports better decision-making and enhances the overall quality of the simulation process.

By implementing these collaborative dynamics, organizations can enhance their engineering simulations, improve data quality, and streamline processes, ultimately leading to faster and more accurate product development.

7. Workflow Automation

Adapting workflows to project needs involves integrating engineering pipelines and automating tasks, enhancing efficiency in resource allocation and scaling.

Automating workflows and accelerating results are essential for managing the growing number of computational analyses involved in engineering and scientific R&D. Connected and automated workflows unify these analyses, orchestrating end-to-end computing processes and data dependencies to increase efficiency, save time, and improve solution accuracy.

End-to-end engineering and scientific process automation connects various multiphysics, AI/ML, and other modeling and simulation workloads into a single executable workflow. By leveraging a comprehensive software library of pre-installed commercial and open-source software, along with the ability to integrate in-house software and scripts for custom pre- and post-processing, organizations can streamline their workflows. This integration facilitates the seamless execution of complex simulations and analyses, ensuring that all necessary tools and resources are readily available.

Adopting and scaling the latest technologies for modern design exploration and product validation is crucial. Workload orchestration accelerates the adoption of AI-augmented solvers and specialized computing

architectures, enabling the confident execution of large-scale workflows with massive parallelization and global resiliency in computing resources. Balancing application performance and cost efficiency within computational pipelines ensures that each workload is optimized for both speed and budget.

Using advanced execution frameworks, engineering and scientific R&D teams can perform large-scale and complex computational pipelines consistently and reliably. Standardizing and scaling modular and flexible computational workflows with configurations and conditional logic ensures that specific data dependencies and hardware requirements are met. This approach provides a robust foundation for executing intricate simulations and analyses, maintaining consistency and reliability across projects.

Facilitating collaboration to meet project requirements is essential for successful R&D. Computational pipelines empower subject matter experts to share requirements and findings at each stage of design exploration and performance optimization. By bringing HPC and domain expertise together with unified workflows, teams can quickly identify optimal design decisions, enhancing the overall efficiency and effectiveness of the R&D process.

Empowering any business user to contribute to R&D use cases with multi-step analyses, from drug discovery pipelines to automotive multi-disciplinary optimization, is key to modern workflow development. Standardizing critical steps for process consistency and traceability while enabling modularity allows teams to flexibly use best-fit tools across various projects. This flexibility ensures that workflows can adapt to the specific needs of different R&D initiatives, promoting innovation and efficiency.

Capturing insights for a complete digital thread is increasingly important as the use of modeling and simulation in science and engineering grows. Automated computational pipelines capture and process data to create a digital thread of findings that can be easily traced and shared. Key performance indicators (KPIs) and relevant context are automatically extracted and tagged, providing reviewers with a holistic and centralized view of project details. This automation enhances traceability and reproducibility, ensuring that all data used and generated in simulations is accurately documented and accessible.

By implementing these workflow automation strategies, organizations can enhance their engineering simulations, improve data quality, and streamline processes, ultimately leading to faster and more accurate product development.

8. Case Studies

Real-world examples illustrate successful transitions, highlighting common challenges and solutions. These cases offer actionable insights into overcoming obstacles during migration.

Case Study: Enhancing Engine Design with Cloud-Based HPC

A pioneering engine company, focused on developing and commercializing an ultra-efficient engine architecture, faced significant challenges in their design process. The company aimed to achieve 30-50% fuel efficiency, which necessitated running numerous simulations, constant iterations, and substantial computational power to meet project timelines. These simulations included analyzing a four-stroke internal combustion engine using computational fluid dynamics (CFD) tools. To fully explore the design space and adhere to strict deadlines, the company needed to conduct injector model characteristic studies, geometrical simplification studies, and reaction mechanism reductions, among other analyses. Running tens to hundreds of models simultaneously

required extensive compute resources and access to a large pool of simulation software licenses, leading to high up-front costs and personnel investments. Even with a powerful internal processing cluster, the ability to run parallel jobs was significantly limited.

By leveraging a cloud-based HPC platform, the company was able to customize their simulation resources for each job. The platform allowed them to dynamically provision HPC resources within five minutes of job submission, run tens of jobs simultaneously without queuing for software or hardware, and execute simulations on-demand using both cloud hardware and CFD software. Results were gathered and delivered to local servers for post-processing and analysis. The platform's secure, intuitive interface, combined with end-to-end data encryption on isolated, encrypted clusters, enabled the company to efficiently execute over 240 simulations, significantly improving their engine design and shortening the design cycle.

Using the cloud-based HPC platform, the company achieved several key benefits. They ran over 240 simulations, revealing previously undiscovered design features. The time to results was reduced by 80% by running jobs in parallel with no wait time for hardware or software. They achieved greater than 73% cost savings compared to using additional internal resources. Jobs were run on an SOC 2 compliant infrastructure, ensuring data security and compliance.

This case study demonstrates the significant advantages of using cloud-based HPC for complex engineering simulations, highlighting the ability to dynamically provision resources, run parallel jobs, and achieve substantial cost savings while maintaining high levels of security and compliance.

Case Study: Accelerating Disease Research with Compute-on-Demand Solutions

A leading genome analysis company, specializing in precision medicine, faced significant challenges in their research efforts. The company needed access to an on-demand HPC solution that enabled customers to flexibly run compute-intensive analyses on samples faster and more cost-effectively. Precision medicine often requires analysis across many biomarkers, and the company's technology allows researchers and clinicians to detect ultra-sensitive and ultra-specific variations in chromosomes. This capability was particularly crucial during the COVID-19 pandemic, where rapid analysis of variant strains was essential for developing targeted antiviral therapies or vaccines.

To address these challenges, the company integrated their genome analysis instrument with on-demand cloud computing capabilities. This integration allowed them to provide a Compute On-Demand service, accelerating up-time, optimizing costs, and simplifying complexity for their customers. The company's focus on complex diseases, including rare diseases and cancer, required substantial computational power to analyze unique genetic signatures that determine disease progression and treatment requirements. Their technology uses optical mapping to detect these signatures by sequencing DNA, which involves extracting long DNA strands, labeling them with fluorescence, and analyzing patterns using nanochannels.

The computational demands of disease and cancer research were ever-increasing, and the company realized they were outpacing Moore's law in their ability to generate and analyze patient data. They needed more computational power to keep up with the growing complexity of genomic events. The company turned to a cloud-based HPC platform to scale their computations. This platform abstracted the research tools to make them user-friendly, ensuring that researchers could focus on their work without needing to learn new cloud technology jargon. By 2020, at least half of the company's customers were using the Compute On-Demand service.

The benefits of this solution were significant. The company achieved 40% faster time-to-answer with improved hardware configurations, reduced sample processing costs by up to 50%, and maintained GDPR compliance with a regionalized compute solution. Personalized clinical treatments, particularly in the fight against cancer, are poised to play a larger role in the future of healthcare. Tools that allow the identification of genetic variants need to be quick and accessible, which on-demand cloud computing facilitates. The company's technology is already making a difference in the lives of those with leukemia, and the potential impact of personalized clinical treatments on the future of medicine is immense.

This case study highlights the advantages of using compute-on-demand solutions for complex disease research, demonstrating how cloud-based HPC can accelerate research, reduce costs, and improve accessibility and performance for critical medical analyses.

9. Conclusion

Transitioning to cloud-based HPC requires strategic planning but offers significant benefits. By addressing challenges with practical strategies, engineers can achieve greater efficiency and cost savings, positioning their organizations at the forefront of innovation.

10. References

- [1] ["Four Ways to Digitally Transform with HPC in the Cloud"](#)
- [2] ["Worldwide HPC in the Cloud Forecast, 2022-2027"](#)
- [3] [Benchmarking Cloud-HPC Hardware for R&D and Engineering Applications](#)

Safeguarding Engineering IP in the Cloud: Strategies for Secure Global Collaboration

Navin Bagga, Msc.(High Performance Computing) Rescale (UK)

Abstract

State-of-the-art high-performance computing (HPC) platforms now integrate a broad range of advanced security capabilities to protect sensitive data and workloads at scale. These include Zero Trust architectures with strong identity verification, fine-grained access control policies, and support for federated identity systems that enable seamless, secure collaboration across organizations. Modern platforms offer encrypted data transfer, policy-based workload orchestration, and secure software environments to ensure supply chain integrity. Compliance with major standards like SOC 2 Compliance, ISO27001, HIPAA, GDPR, ITAR and FedRAMP is increasingly standard, alongside features like auditability and centralized governance that support both operational transparency and regulatory alignment.

As engineering and simulation workloads shift to cloud environments, securing sensitive intellectual property (IP) has become a critical priority for all organisations. This trend introduces unique challenges, particularly when enabling multiple organisations and cross-continent collaborations. However, advancements in cloud computing and cybersecurity provide robust strategies to protect valuable data while maintaining operational efficiency. This presentation explores these strategies, focusing on real-world use case needs when accessing cloud based HPC and best practices for safeguarding engineering IP in public cloud environments.

A key focus will be on specifying and managing compute and storage preferences to achieve optimal performance, cost-efficiency, and security. The session will delve into methods for managing end-user permissions, ensuring compliance and minimising risks. Attendees will also gain insights into the benefits of running simulations and engineering workloads in their preferred cloud regions while maintaining stringent data protection protocols.

The presentation will highlight how to ensure Data Security and Compliance of HPC simulation data, when

running in the Cloud, a pivotal tool for organisations aiming to retain complete control over their data. Additionally, the discussion will cover strategies for managing data flows and transfers in distributed environments using a Zero Trust Security framework. Attendees will learn how leading enterprises are addressing challenges like data sovereignty and compliance through innovative security measures. The presentation will also explore how these practices enable enterprises to safeguard sensitive engineering workflows, facilitate secure global collaborations, and prepare data for artificial intelligence and machine learning applications.

This session is designed for IT leaders, engineers, and cybersecurity professionals seeking actionable insights into cloud-based engineering solutions. By combining cutting-edge technologies and practical examples, this presentation will empower attendees to enhance their cybersecurity frameworks and optimise their cloud engineering operations while ensuring the confidentiality and integrity of their IP.

1. Introduction

As HPC (High Performance Computing) infrastructures grow in demand and serve the ever increasing growth in HPC consumption of organisations, the complexity and access from various departments that need to tap this unlimited compute & data resource is inevitable. An organisation therefore is placed with greater responsibility on choosing the right Security and Compliance initiatives that cover various aspects of HPC scenarios. This includes cases such as data sovereignty to be respected in its various forms, data leakage to be restricted and concerns with confidentiality whilst collaborating simulation data, to name a few.

The security standards and frameworks such as NIST / SOC Attestation / GDPR and ISO 27001 compliance provide an organisation with great confidence on the underlying HPC infrastructure components. However, in a shared security model, it is still the responsibility of an organisation utilising cloud resources to ensure their side of compliance is adhered to the regional, industry and the project based mandates. Organisations like Firefly Aerospace, Kairos Power and Virgin Orbit require highly confidential compute environments hosted in isolated access environments and are yet able to access the power of cloud with security first principles as described in this paper.

Concepts like Zero trust security are no longer an ultimate goal, but an essential part of a scalable HPC infrastructure, especially when conducting engineering research and developing intellectually guarded outputs which lie at the heart of any organisation's Intellectual Property (IP).

In this paper we shall aim to cover the needs from the HPC community related to the security and compliance posture management with a focus on "Simulation Data collaboration" and how Rescale's HPC platform helps to address such scenarios.

2. Security in the HPC Industry

HPC industry is no stranger to evolving security needs that go hand in hand with evolving HPC architectures, solver suites, applications, libraries and associated dependency chains. Security is part of every step when working on simulation data, right from the initial entry to a system, to authentication, SSL (Secure Sockets Layer) encryption in various forms for client side and data encryption whether the data is 'at-rest' or 'in-transit', to cluster

privacy and safeguarding the simulation data output.

We have observed an increase in the focus on security, especially with the growth in demands for Cloud burst capacities for HPC infrastructures as well as SaaS/PaaS based services to achieve the economies of scale that most modern enterprises expect. To stay current, there is always an ever growing list of Security and Compliance accreditations that any solution must adhere to. This places quite a hefty effort on organisations that want to explore the cloud capabilities but are inundated with allocating resources for such growing demands. This is where there is a genuine need for solutions to provide ‘Security as a Service’ on top of the underlying capabilities being offered to enhance an HPC infrastructure.

Rescale holds security leadership in cloud simulations by not just adhering to these needs but by going beyond the expected to provide its customers with administrative controls, strategic guidance, tailored security and compliance support to achieve the evolving security needs of simulation data collaboration.

3. Rescale’s stance on Secure HPC in the Cloud

At Rescale, ‘Zero Trust’ is a comprehensive methodology that encompasses security architecture, security policies and security procedures. Rescale’s defense-in-depth security strategy focuses on four key security facets: people, devices, networks, and workloads. ‘Zero Trust’ means that trust in any one of the four key areas does not implicitly mean trust in another. Any entity (people and devices) requires authenticated trust at every level. Rescale constantly monitors all four areas to ensure that entities are operating securely within Rescale’s trust model.

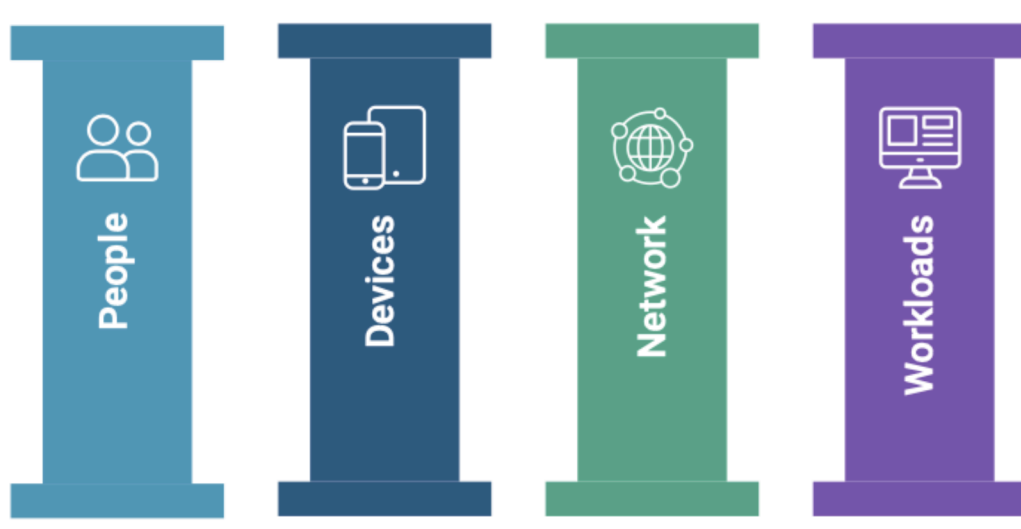


Figure 3.1 Security pillars of a modern HPC Service

Key outcomes for achieving Secure HPC in the Cloud involves ensuring:

- Only authorized individuals (people) have access to simulation data
- Authorized users are accessing company and platform data from secure endpoints (devices).
- All network access must be authenticated at all times, and to limit the flow of data to specific ports and protocols as defined by the principle of least privilege.
- Workloads are isolated from each other, are only allowed to move data within specified ports and protocols as defined by the principle least privilege, and require authentication to access at all times.

Rescale adheres to the cybersecurity framework established by the National Institute of Standards and Technology in the U.S. Department of Commerce. The NIST cybersecurity framework guides Rescale in following all modern security best practices to mitigate risk and secure data and systems. Rescale also uses independent third-party (external) auditors to validate and verify the security posture and maintain the current trifecta of cybersecurity certifications. Rescale currently holds the following certifications as minimum operating standards: American Institute of Certified Public Accountants (AICPA) SOC2 Type2 attestation, US International Traffic in Arms Regulations (ITAR) Compliance, Cloud Security Alliance (CSA), Health Insurance Portability and Accountability Act (HIPAA) compliance, Federal Risk and Authorization Management Program (FedRAMP) Authorised, GDPR Compliance, ISO 27001 compliance and TISAX Level 1 compliance. Rescale also follows the NIST 800-53 framework.

4. Collaboration and Safeguarding Engineering IP

Secure Model Based Collaboration

As organizations build up their computational pipelines and better integrate their supply chains, seamless and secure collaboration is an essential element to support best-in-class R&D amongst teams working on multiple projects.

But managing access and policies for shared high performance computing (HPC) resources (hardware and software) and the related simulations, can be massively complex while also opening up your organization to data exposure and other risks to intellectual property. Not only do the HPC controls need to cater to the demands of security, but there is a continuous need to collaborate amongst teams and team members. This brings in various challenges such as 'Regional Compliance', 'Data Leakage Avoidance' and 'Simulation Data Privacy' controls. With the growing demand of Artificial Intelligence (AI) based workloads, these needs are catapulted to another level. Where there were users and teams alone, now will be agents, bots and service identities added to the mix.

Rescale provides a centralized environment for any collection of users that need to collaborate closely and share the same HPC resources, regardless if they are in the same organization or collaborating from another organization (partner in the supply chain) with full support for the evolving automation needs due to the growing demand in AI based workflows.

Following are the key aspects that form the basis of an HPC service that can offer secure model based collaboration:

Fine-Grained Access Control

The HPC service must offer comprehensive visibility with policy-based financial and architectural controls, allowing organizations to define and enforce detailed access policies for various groups that will need to access HPC.

Federated Identity & Cross-Site Authentication

The HPC service must support features such Single Sign-On (SSO) integration, MFA (Multi-Factor Authentication) and facilitating federated identity management with streamlined cross-site authentication processes.

Container and Software Supply Chain Security

The HPC service must provide software management capabilities, enabling access to licensed, open-source, and custom applications instantly, which aids in maintaining software supply chain integrity.

Network Segmentation & Encryption

The HPC service must ensure full-stack encryption and authentication controls, safeguarding data in transit and at rest.

Secure Job Scheduling

The service must include workload orchestration features, allowing for secure and efficient job scheduling across various computing resources.

Auditability and Compliance

In the Zero Trust model, the underlying system must strictly capture auditing of information from all the pillars of a modern HPC service, such as users, devices, networks and workloads; as shown in the previous section. Auditing of User behaviour, in particular, can be challenging as the system must capture the relevant information without compromising the local laws such as GDPR compliance. The HPC service or system must be capable of providing information that can then be analysed to make meaningful analysis to trace back an action in a given audit. Such auditing of information may be enriched by the user location, type of device, even down to the type of application or browser utilised with their specific version(s), if possible.

This forms the basis for a service that can be utilised for wider collaboration of HPC and Engineering IP. Now, let's look at the collaboration features and models to employ when accessing HPC for various personas.

Regional Controls

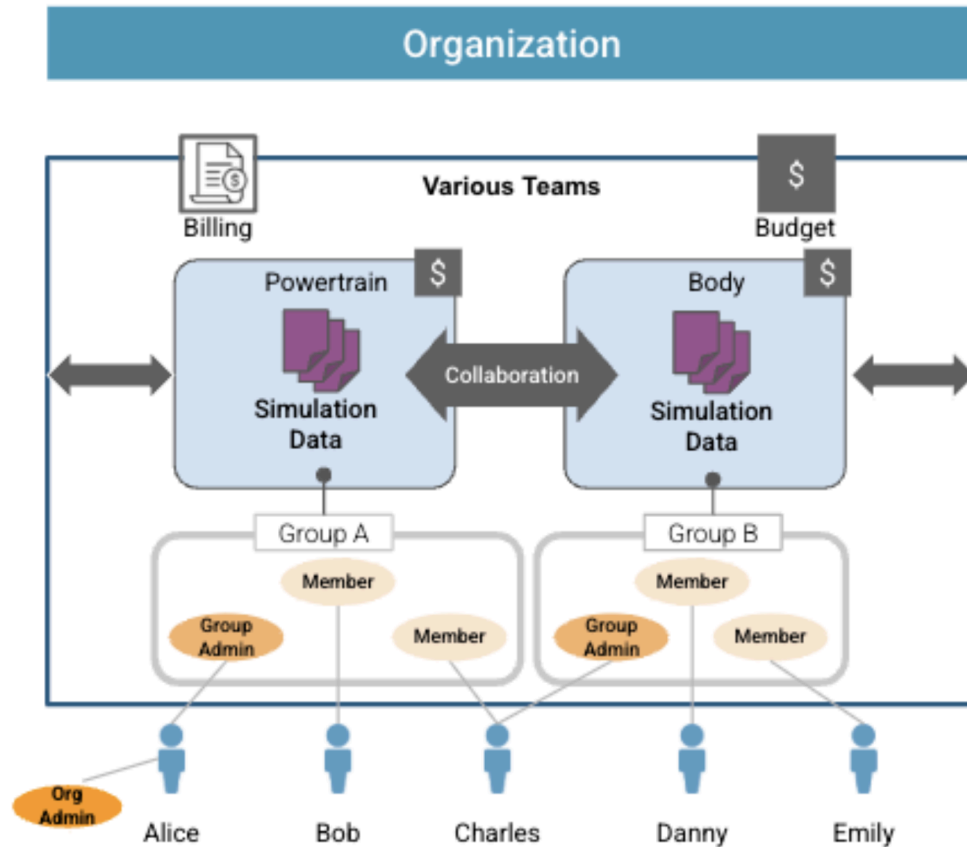
In many cases, organizations must maintain regional boundaries, particularly due to local compliance laws. Often, collaboration happens only under certain circumstances by project. Other times, teams from different supply chain partners need to work together continuously but should only have access to certain resources and data. The HPC system should cater to Regional compliance for compute and data whilst ensuring the controls do not limit the choice and scalability of the underlying Cloud infrastructure.

Rescale provides its customers the ability to restrict Compute and Data Regions by applying Regional Preferences to the Customer's Workspace(s). This means, a Customer can choose to operate in certain Regions and Data Center locations of their choice, irrespective of the underlying Cloud Service Provider(CSP). This allows a customer to be strictly aligned with their Regional compute and data needs without compromising on the scalability or availability of Cloud computing resources.

Tightly Coupled Collaboration

There are times when an organisation must cater to a collection of users that need to collaborate closely and share the same HPC resources, regardless if they are in the same organization or collaborating from another organization (partner in the supply chain, for example). Despite multiple engineering teams working in various regions, the organisation must endure capabilities to create a single, central environment where anyone with approved access can collaborate.

Critically, this need is to ensure the prevention of the manual cycle of copying and pasting files from local drives to share, which introduces risks of Data Leakage of Engineering IP. The more teams carry out these types of unmanaged tasks, the more is the risk of data leakage.



Tightly Coupled Collaboration - High level of trust

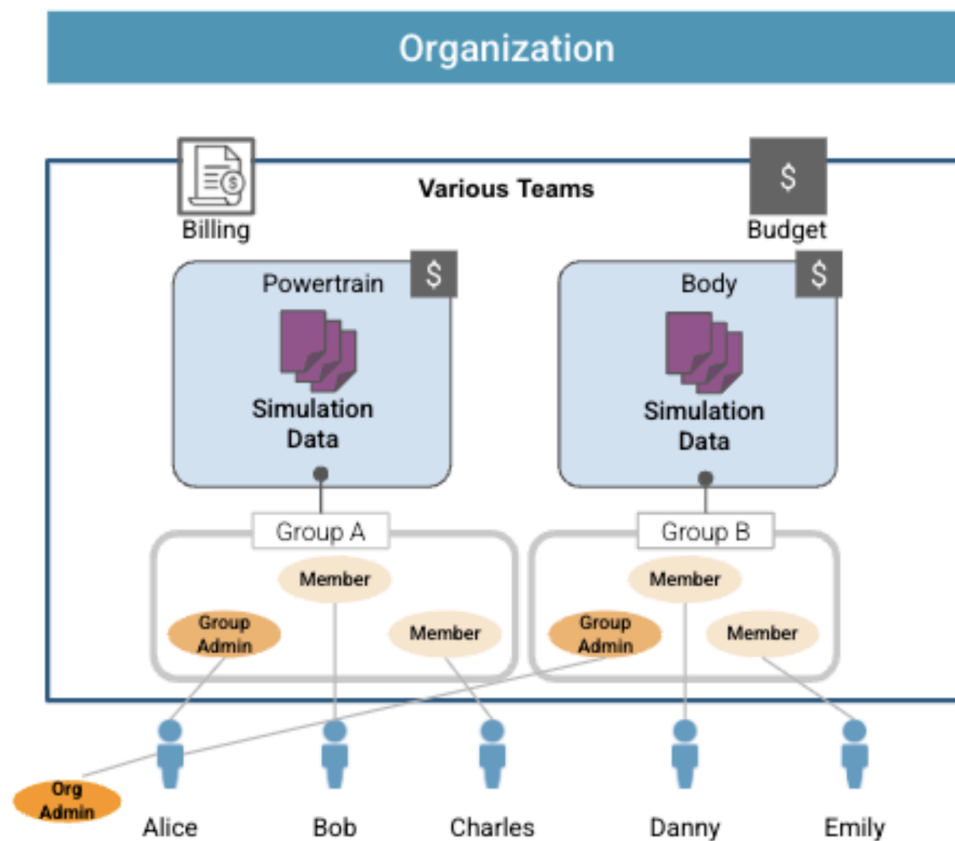
Figure 4.1 Example of an Organization with Various collaborating groups

In this scenario, the level of trust among all the groups in the organization is high. Rescale’s out of box solution with a single Workspace provides a simple structure for bringing the digital engineering teams together. While the Workspace provides the central environment for all engineers to access the same HPC resources, the organization is using the platform to also manage work at the project level, which allows engineering teams to share jobs, workstations, budgets, data, and simulations with just the people assigned to those projects within the Workspace. This provides a centralized control, without the need to exfiltrate the engineering and simulation data from the defined Regions, Groups and/or Projects. Applying the principle of least privilege, Rescale customers are able to implement a fully global collaborative solution whilst still maintaining centralised compliance and regional control.

Loosely Coupled Collaboration

On another end of the collaboration spectrum, is a need for the shared usage of resources in an HPC infrastructure with full security: complete isolation of each user and no visibility into the simulations for any user. We call this “loosely coupled” collaboration, which is really about ensuring highly restricted access to HPC resources on demand.

Critically, the organisation doesn't want its users to have any visibility or access to other users and their data, just like on a typical public cloud service or SaaS platform. The end users may upload their simulation data via the web interface which is encrypted in transit, submit simulation jobs to run on cloud services with encrypted cluster disks and get the results back via the web interface following the same data security protocols. Behind the scenes, each user's access is tied to a unique API token only known to the end user. This ensures that every transaction is cryptographically sealed to keep each user isolated with access only to specific Rescale resources.



Loosely Coupled Collaboration - User isolated roles

Figure 4.2 Example of an Organization with Various isolated groups

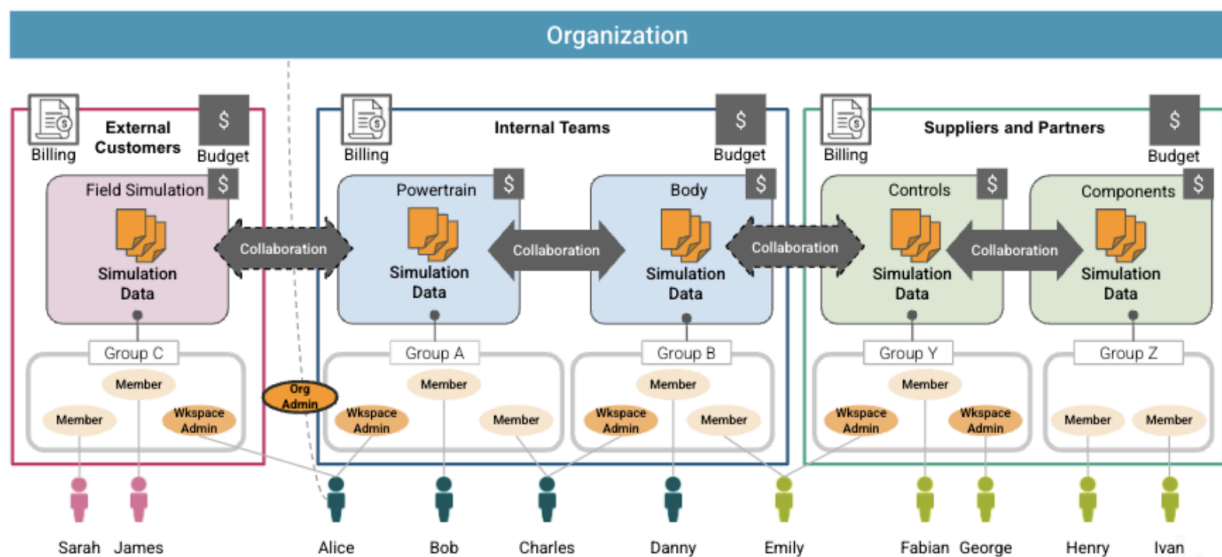
For an HPC infrastructure that demands high levels of user isolation, it is additionally possible for Rescale customers to develop a frontend that talks to the Rescale platform for on demand HPC infrastructure provisioning. The users don't even know Rescale is working in the background or that it hosts the simulation. Rescale policies easily maintain absolute security segmentation among all the Users that log in to the platform

to access and run simulations. Data that is unique to each user is in no way accessible by another user.

Blended Collaboration

In many organisations, HPC collaboration requirements are not an all or nothing situation for access to data, regional preferences and simulation jobs. This is not uncommon that many organisations need to apply a mix of collaboration approaches on their HPC solution to control sharing and permissions, depending on various relationships between groups.

A blended approach can be ideal for close collaboration between supply chain partners, for example. There is a certain degree of trust, but also a degree of caution and control to keep various entities separate while also enforcing the proper administrative demarcation, such as with the use of regional preferences for the underlying cloud services.



Blended Collaboration - Role based segregation and collaboration

Figure 4.3 Example of an Organization with Various internal and external groups

5. Conclusion and Future Directions

In this paper, we looked at what Security and Compliance challenges are faced by organisations who need to collaborate in various forms in regards to their simulation data whilst strictly adhering to Regional controls, Security mandates and protecting their IP when using cloud based HPC resources.

To cater to and service the ever evolving ‘Security & Compliance’ requirements, organisations need to think of their toolsets in terms of automations and ready to use services. A DIY (do-it-yourself) approach for HPC Security can involve a significant amount of effort and resources for an organisation, as such, it’s worth considering a tailored solution that can provide a ‘Security as a Service’ model. Such an approach enables an organisation to collaborate securely without a long implementation time for various collaboration scenarios, as discussed in this abstract. This also ensures that the efforts of a given organisation are best focussed on developing their engineering and research outcomes whilst partnering to meet the ever evolving Security and Compliance requirements.

We looked at how Rescale’s Secure HPC as a Service platform provides an out of box solution to such scenarios where an organisation needs to cater to complex global, local and individual security restrictions when handling engineering IP.

6. References

- [1] [The Impact of Model-Based Systems Engineering & Collaboration](#)
- [2] [Cloud Computing Security | Rescale](#)
- [3] [Leading HPC Security and Compliance Built for the Cloud - Rescale](#)
- [4] [Rescale Zero Trust Security PDF](#)
- [5] [Rescale Powers Seamless and Secure Simulation Collaboration for Global Organizations and Supply Chains](#)
- [6] [Model-Based Collaboration with Rescale Metadata Management](#)
- [7] [Customers - Rescale](#)

Orchestrating Hybrid HPC Environments: Strategies for Data Gravity and AI-Ready Datasets

Romain Klein, Rescale (France)

Carlos Mecha, Rescale (USA)

Abstract

High-performance computing (HPC) continues to play a pivotal role in driving innovation across industries, from design and simulation to AI-enabled engineering. As workloads increasingly shift between on-premises and cloud environments, orchestrating hybrid HPC systems has become essential to ensure performance, cost-efficiency, and scalability. This presentation will delve into strategies for addressing data gravity challenges and implementing robust data management solutions in hybrid HPC architectures.

A central focus will be on the evolving role of data strategies in HPC environments. The paper will cover the importance of tiered storage strategies, highlighting how balancing fast-access storage tiers with cost-effective archival solutions can optimise operational costs without sacrificing performance. Emphasis will be placed on the critical role of Simulation Process and Data Management (SPDM) and Simulation Process and Data Resource Management (SPDRM) systems in ensuring data provenance, reproducibility, and accessibility for downstream use cases.

One of the key innovations to be discussed is the Data Lake Exporter (DLE) solution. This approach automates the transfer of HPC workloads executed in the cloud to archival cold storage, enabling organisations to minimise costs while preserving simulation results for future use. Furthermore, this process begins the transformation of raw simulation data into AI-ready datasets, opening pathways to leverage machine learning and AI for predictive modelling, optimisation, and insight generation.

The presentation will also explore best practices for metadata capture to ensure simulation data is not only stored but also structured for effective retrieval and AI utilisation. Real-world use cases and insights from deploying hybrid HPC solutions across diverse industries will provide context for the strategies shared.

Attendees will leave with a deeper understanding of how to design and manage hybrid HPC environments that seamlessly integrate compute, data management, and AI workflows while addressing sustainability goals and operational efficiency. This session will be particularly valuable for engineering leaders, data scientists, and HPC practitioners aiming to harness the full potential of hybrid architectures in their simulation and AI endeavours.

1. Introduction

HPC systems are the backbone of advanced engineering simulations, computational research, and AI workloads. Traditionally, these systems have been deployed in on-premises clusters, but the increasing complexity and computational demands have driven organizations toward hybrid cloud solutions. Hybrid HPC environments combine the scalability of the cloud with the control and performance of on-premises infrastructure, creating new challenges in data management, storage, and AI integration.

This paper presents best practices for tiered storage strategies, discusses the significance of Simulation Process and Data Management (SPDM) and Simulation Process and Data Resource Management (SPDRM) systems, and introduces the Data Lake Exporter (DLE) as a solution for long-term data archiving and AI-readiness. Real-world applications will be examined to provide actionable insights for engineering leaders and HPC practitioners.

2. Data Gravity Challenges in Hybrid HPC

The rapid adoption of hybrid and distributed high-performance computing (HPC) architectures; spanning on-premises, colocation facilities, and public/private cloud clusters, has led to an explosion of data stored redundantly across multiple systems. This uncontrolled proliferation of data creates data gravity challenges, where large datasets become increasingly difficult and expensive to move between environments.

Growing Data Explosion from Increased Simulation in R&D

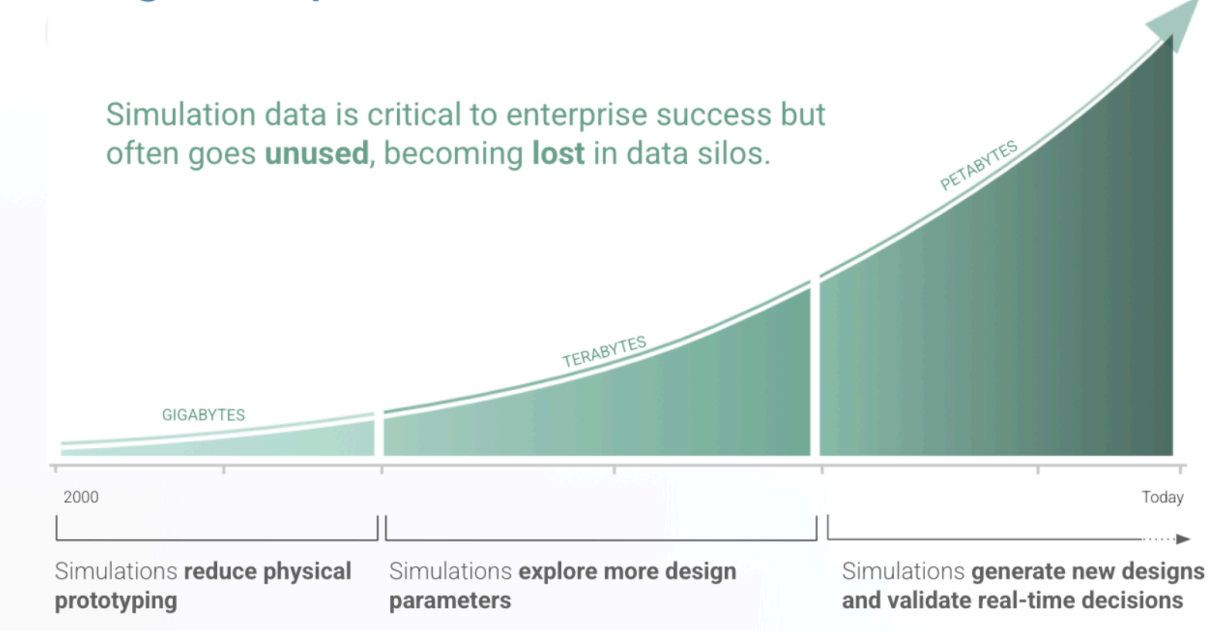


Figure 1: ExGrowing Data plosion from Increased Simulation in R&D

The Impact of Data Gravity on Hybrid HPC

Data gravity arises when massive datasets accumulate in disparate storage locations, making their movement impractical due to:

- **High Transfer Costs:** Moving terabytes or petabytes of data between clusters or between on-premises and cloud environments incurs substantial egress fees.
- **Time-Consuming Transfers:** Limited network bandwidth can result in data movement taking hours, days or even weeks, causing workflow bottlenecks.
- **Lack of a Single Source of Truth:** Storing multiple copies of datasets across different clusters leads to inconsistencies, making it difficult to track the latest version.
- **Data Provenance and Reproducibility Issues:** The absence of centralized version control for simulation data complicates efforts to validate, reproduce, and audit CAE workflows.

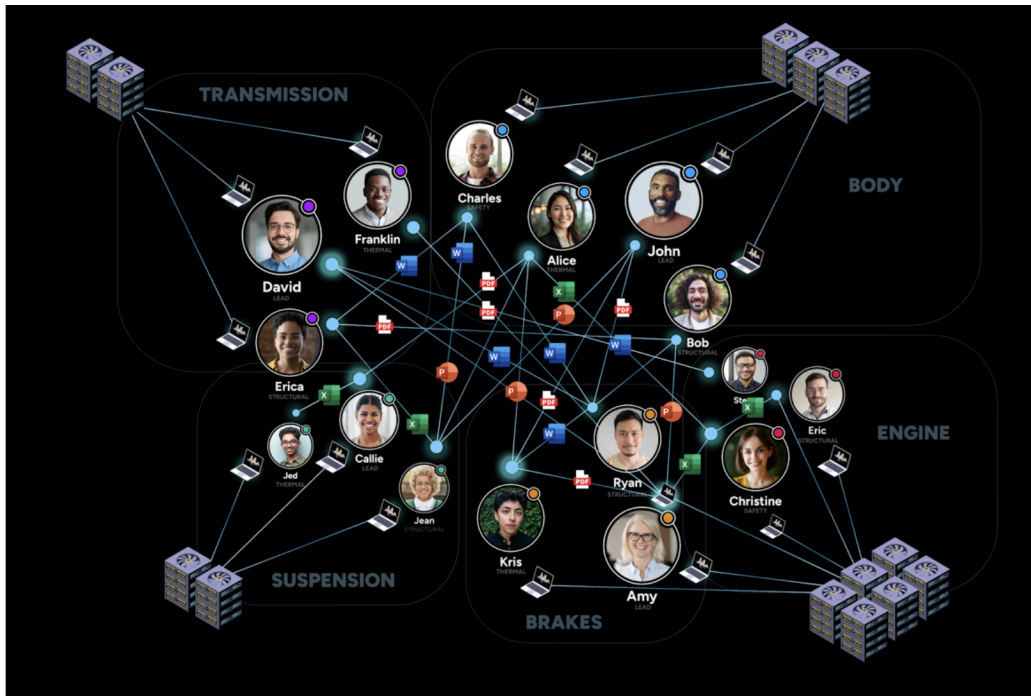


Figure 2: R&D spans a large number of teams and disciplines collaborating on simulation data

Mitigation Strategies for Data Gravity Challenges

To address these challenges, organizations should adopt proactive data lifecycle management and structured data governance strategies:

a) Proactive Data Lifecycle Management

- **Classify and Tag Data Automatically:** Integrate Simulation Process and Data Management (SPDM) or Simulation Process and Data Resource Management (SPDRM) systems with hybrid HPC clusters to track simulation inputs, outputs, and metadata.
- **Eliminate Redundant Copies:** Implement deduplication policies to avoid unnecessary replication of simulation data across multiple storage locations.
- **Prioritize Local Storage for Active Workloads:** Frequently accessed data should reside in high-performance local scratch storage, while completed workloads should be archived in cost-efficient cold storage.

b) Storage Tier Strategies for On-Premises and Cloud Data Centers

A well-defined tiered storage strategy is crucial to balancing performance and cost efficiency across hybrid HPC environments.

Storage Tier	On-Premises HPC	Cloud HPC	Use Case
Tier 1 (High-Performance Storage)	Local NVMe, Lustre, BeeGFS	AWS FSx for Lustre, Azure Managed Lustre	Active simulations requiring high-speed I/O
Tier 2 (Warm Storage)	NAS, Object Storage	Cloud object storage (S3, Azure Blob)	Intermediate data access, checkpointing
Tier 3 (Cold Storage)	Tape Libraries, Archival NAS	AWS Glacier, Azure Blob Archive	Long-term storage, compliance archiving

Table 1: Storage Tier Strategies for On-Premises and Cloud Data Centers

c) Optimizing Data Transfer for Time and Cost Efficiency

Strategically placing data near the most optimized compute resources reduces latency and costs associated with data transfers. This can be achieved through:

- **Workload-Aware Data Placement:**
 - AI and simulation workloads that require rapid access should be co-located with HPC clusters that provide the best compute-to-storage ratio.
 - Batch workloads can be scheduled in locations where compute and storage costs are lower.
- **Region-Based Data Optimization:**
 - Geo-distributed HPC clusters should route data to the nearest available compute node to minimize

transfer delays.

- Metadata feeds from SPDM/SPDRM systems can automate data locality decisions.
- **Department-Specific Storage Policies:**
 - Engineering teams can define department-level retention policies to keep datasets only where they are needed, avoiding unnecessary cloud transfers.
 - Cross-department collaboration can be enhanced by providing controlled access to centralized repositories rather than duplicating files.

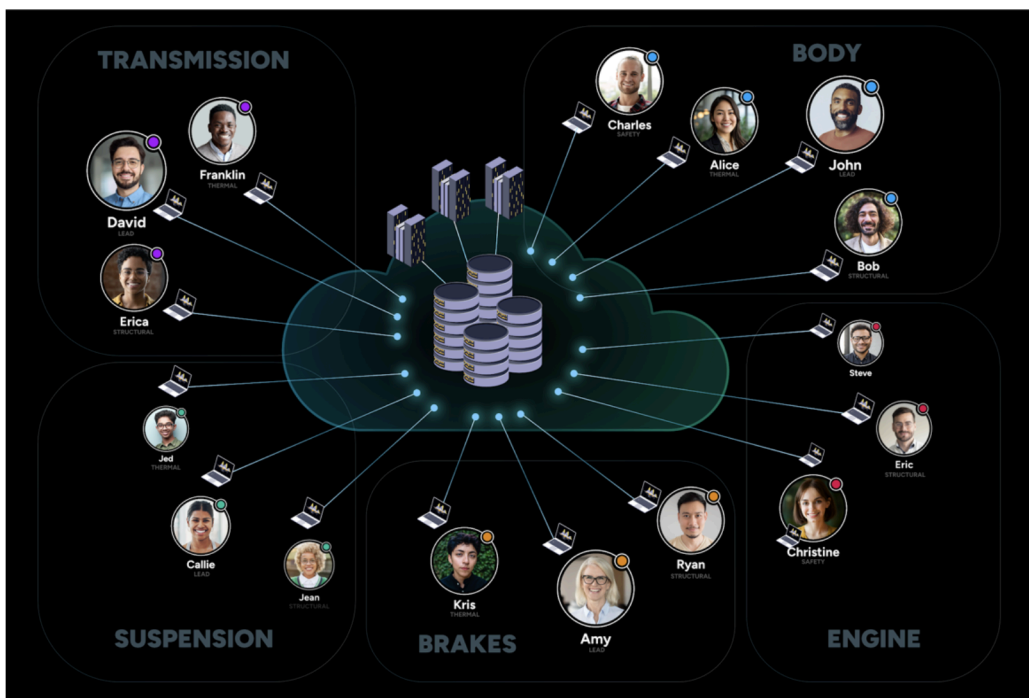


Figure 3: Unified digital thread that connects every piece of data, every decision, across all teams and disciplines.

3. Solutions for Addressing Data Gravity Challenges

As hybrid HPC environments grow in complexity, overcoming data gravity requires intelligent data management solutions. Two critical approaches to mitigating these challenges are Data Lake Exporter (DLE) for Efficient Archiving and Automations for Data Classification and Tagging.

A. Data Lake Exporter (DLE) for Efficient Data Archiving

Another key strategy for managing data gravity is efficiently archiving large datasets. The Data Lake Exporter (DLE) ensures that simulation data is systematically transferred to long-term, cost-efficient storage, avoiding unnecessary duplication and freeing up high-performance storage for active workloads.

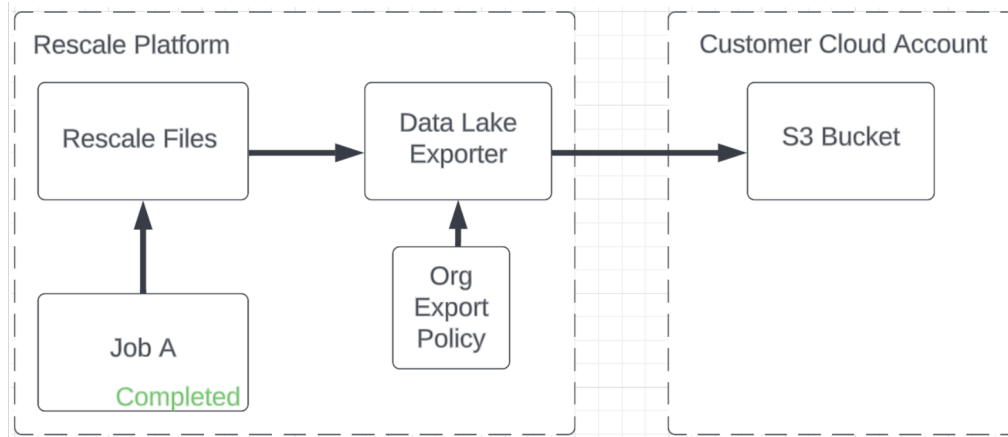


Figure 4: Example implementation of the Data Lake Exporter

For organizations using DLE, the system automatically scans and exports job-related files based on predefined policy filters. The process works as follows:

- 1. Job Completion Detection:** DLE monitors completed jobs and checks them against predefined policy filters (e.g., project type, owner, software version).
- 2. File Filtering:** A second layer of filtering is applied using regular expressions on file names to identify relevant data.
- 3. Data Transfer to Object Store:** The selected files are transferred to a designated cloud storage bucket (e.g., AWS S3) for the organization. Data is structured within folders labelled with job IDs for easy retrieval.

Use Case: Automating Data Archival

A hybrid HPC organization using Rescale and AWS S3 can implement DLE to:

- Automatically export completed simulations to a low-cost data lake for future analysis.
- Free up high-performance storage for active workloads, optimizing resource utilization.
- Maintain traceability and compliance by structuring archived data in an organized, searchable format.

Filter Type	Example Filters
Job Owner	Specific users or teams based on email addresses.
Software Used	Specific CAE software (e.g., Ansys, Abaqus).
Software Version	Restrict data syncs to specific versions for consistency.
File Type Filters	Regex-based filtering to include only essential job files.

Table 2: The DLE syncs a filtered subset of job-related files based on organization-specific configurations

The Value of DLE for Hybrid HPC Workloads

- **Cost Savings:** Shifts inactive data to cost-efficient archival storage, reducing cloud storage expenses.
- **Performance Optimization:** Frees up high-speed storage for active simulations.
- **AI Readiness:** Prepares archived datasets for machine learning and predictive analytics.

B. Automations for Automatic Data Classification and Tagging

One of the most effective ways to reduce data gravity is to automate the classification, tagging, and movement of data throughout the simulation lifecycle. By integrating pre-, during-, and post-job automations, organizations can reduce manual interventions and ensure that data is properly categorized, accessible, and efficiently managed.

Key Features of Automation Solutions

- Pre-, During-, and Post-Job Automations
 - Configure workflows to classify and tag data at different simulation stages, ensuring real-time organization.
 - Reduce the risk of misplaced, untracked, or redundant data across hybrid environments.
- Marketplace Integration
 - Leverage a library of pre-built and user-generated automations to streamline data processes.
 - Accelerate project timelines by using existing workflows rather than building custom solutions from scratch.
- Developer-Friendly Tools
 - Use Python or Bash scripting to create and deploy automations tailored to specific simulation requirements.
 - Enable deeper customization through integration with external APIs.
 - Flexible Customization
 - Define key-value parameters to dynamically modify automation behaviour.
 - Adapt automations to specific team requirements while maintaining standardization across workflows.
- Workspace-Level Controls
 - Administrators can enforce consistent automation policies across teams to maintain governance and best practices.
 - Reduce operational inefficiencies caused by disparate manual approaches to data classification.

Automation Type	Use Case
Metadata Extraction	Automatically extract and tag simulation parameters (e.g., peak stress, max temperature) into metadata fields for easy searchability
Slack/Microsoft Teams Notifications	Send real-time job status updates, completion alerts, or error notifications to communication channels.
File Management	Automate file compression, format conversions, and storage transfers to cloud services like AWS S3.
Pre-Processing Checks	Validate simulation inputs, estimate job runtime, and flag errors before execution to prevent unnecessary failures.
Environment Customization	Automatically apply configurations such as workstation setup, network drive mapping, and software extensions.
LLM Integration	Use AI models like OpenAI to generate simulation summaries and analyze job logs for optimization insights.

Tale 3: Example Automations in HPC Workflows

Real-World Impact: Transforming HPC Workflows

Traditional storage systems capture raw files, but automations transform them into actionable intelligence. Engineers benefit from immediate insights, reduced manual work, and faster decision-making.

For example, metadata extraction allows simulation inputs and key findings to be automatically enriched with contextual information, forming a "digital thread" across projects. This eliminates the need for engineers to manually sift through datasets, accelerating analysis workflows.

4. Conclusion

Managing data gravity in hybrid HPC environments requires automated data classification, structured archiving, and intelligent storage tiering. Data Lake Exporter enables seamless data export to cost-effective archival storage, while Automations simplify metadata tagging and workflow management. Together, these solutions help organizations reduce costs, improve performance, and enhance AI-driven insights.

5. References

Digital Thread Foundations for Accelerated Multi-Disciplinary CAE Workflows

Madhu Vellakal, Rescale (USA)

Garrett VanLee, Rescale (USA)

Abstract

The growth of computer-aided engineering (CAE) tools and simulation data offers new opportunities for engineering teams to develop new products faster. However, challenges persist due to fragmented workflows, siloed simulation data management, and inefficient manual metadata processes. Engineers often spend up to most of their time managing data [1] instead of making critical design decisions, impacting time-to-market, operational costs, and the ability to deliver optimal solutions.

Despite its many advantages, the use of CAE in the product development process presents several challenges. One major hurdle is the high computational power and specialized software required for advanced simulations. Additionally, accurately modeling real-world conditions in a virtual environment is complex and may not always capture all the nuances of physical behavior, leading to discrepancies between simulation results and real-world performance. CAE also relies heavily on accurate material data and boundary conditions; any errors or assumptions in these inputs can lead to inaccurate predictions. Lastly, while CAE speeds up development by reducing the need for physical prototypes, it may still require validation through physical testing, which can introduce delays and costs. These challenges highlight the need for continuous improvement in CAE technologies and the expertise of engineers to fully harness its potential.

In this abstract, we will talk about how a Graph Neural Network (GNN) architecture is deployed for modeling the complex behavior of bipolar plates of PEM (Proton Exchange Membrane) fuel cells. Modeling fuel cells involves complex FEA and CFD methods. Geometry preparation for the FEA process is human intensive and solving the FEA simulation takes a minimum of 48 hours on 100s of CPUs. The deformed geometry from the FEA simulation is processed into a CFD model for the flow prediction. Using a surrogate model approach we will demonstrate how we can predict the structural deformation of the geometry starting from the CAD model. Inferencing on new designs with the surrogate model reduces design validation time from 48 hours with traditional methods to just seconds in real-time.

A framework to streamline multidisciplinary simulation workflows by integrating digital thread concepts and AI-driven methodologies is discussed. By unifying historical modeling and simulation data, automating metadata capture, and leveraging AI for optimization, this framework approach significantly enhances collaboration,

decision-making, and productivity.

1. Introduction

CAE offers numerous benefits, but its application in product development comes with a few challenges. One key obstacle is the significant computational power and specialized software needed for advanced simulations. Moreover, accurately replicating real-world conditions in a virtual setting is complex and may not fully capture all the subtleties of physical behavior, resulting in discrepancies between simulation predictions and actual performance. CAE also depends on precise material data and boundary conditions, and any errors or assumptions in these inputs can lead to flawed predictions. While CAE accelerates development by reducing the need for physical prototypes, it often still requires validation through physical testing, which can introduce delays and additional costs. These challenges underscore the importance of continuous advancements in CAE technologies and the expertise of engineers to fully leverage its capabilities.

The framework proposed addresses key engineering challenges by introducing centralized data structures, automated processes, and AI-assisted modeling. A digital thread tracks the entire lifecycle of simulation workflows by integrating data and metadata from project details directly into engineering simulations. It ensures accessibility for collaborators while maintaining a comprehensive record of all decisions made and actions taken across related and historical simulation activities, providing traceable and actionable insights across teams and disciplines. This ensures that data is not only accessible but also actionable, enabling engineers and decision-makers to make informed choices that accelerate development and improve product outcomes.

Key topics this paper discusses include:

- **Engineering Challenge:** Managing multidisciplinary CAE workflows is increasingly complex due to disconnected tools, fragmented datasets, and manual processes. This challenge is amplified by the explosion of engineering data—both in variety and volume—reaching petabyte scales. In industries like automotive and aerospace, where thermal, structural, and aerodynamic simulations must seamlessly converge, underutilized data leads to duplicate work, difficulty in searching and retrieving information, lost insights, and overall engineering inefficiency.
- **Digital Thread:** A centralized simulation data management system provides a structured, hierarchical framework (e.g., project > folder > study > job) to efficiently organize and share simulation files, findings, and metadata. These systems align simulation activities with business and project objectives through goal-driven project management. Advanced data analysis and exploration capabilities enable teams to compare results, conduct meta-analyses, and uncover new insights. Integrated collaboration tools, such as real-time results sharing and commenting, streamline hand-offs, enhance data accessibility, and improve decision-making. When coupled with AI models, these systems drive predictive insights, accelerate simulations, and optimize iterative design processes.
- **AI-Driven Case Study:** This paper presents a case study from an innovative fuel cell manufacturer, showcasing tangible benefits such as a 50% reduction in simulation runtimes, increased simulation throughput, and improved collaboration across global engineering teams. These results highlight the power of AI and digital thread technology in overcoming bottlenecks, optimizing resource utilization, and driving innovation.

- **Business Outcomes:** The implementation of these methodologies has led to business impact, such as faster product development, reduced operational costs, enhanced sustainability, and higher-quality designs. By enabling real-time collaboration and decision-making, these solutions empower teams to meet customer demands more effectively.

Artificial intelligence (AI) is emerging as a transformative tool in addressing the challenges encountered in running CAE simulations for various applications. AI-powered surrogate models can efficiently approximate complex CAE simulations, enabling rapid evaluations of design alternatives and scenario analyses. Moreover, AI techniques, such as deep learning, facilitate data-driven approaches for enhancing simulation accuracy, predicting aerodynamic behaviors, and identifying critical flow phenomena. Additionally, AI-driven optimization algorithms can efficiently search vast design spaces to identify optimal configurations and performance-enhancing parameters for defense systems. By leveraging AI technologies, researchers and engineers can overcome computational bottlenecks, expedite design iterations, and unlock new insights into aerodynamic phenomena, thereby advancing the development of next-generation products with improved performance, efficiency, and mission effectiveness.

Graph neural networks (GNNs) architectures have emerged as a powerful approach for handling complex data structures and are finding increasing utility in CAE simulations. By representing the computational domain as a graph, with nodes corresponding to grid points and edges denoting connections, GNNs can capture intricate relationships and dependencies inherent in complex physical phenomena. This representation allows GNNs to effectively model spatial interactions, boundary conditions, and turbulence effects, leading to more accurate and efficient simulations. GNNs offer the ability to learn from large-scale datasets, enabling data-driven approaches for optimizing simulation parameters, predicting flow behavior, and identifying critical flow features.

2. Engineering Problem

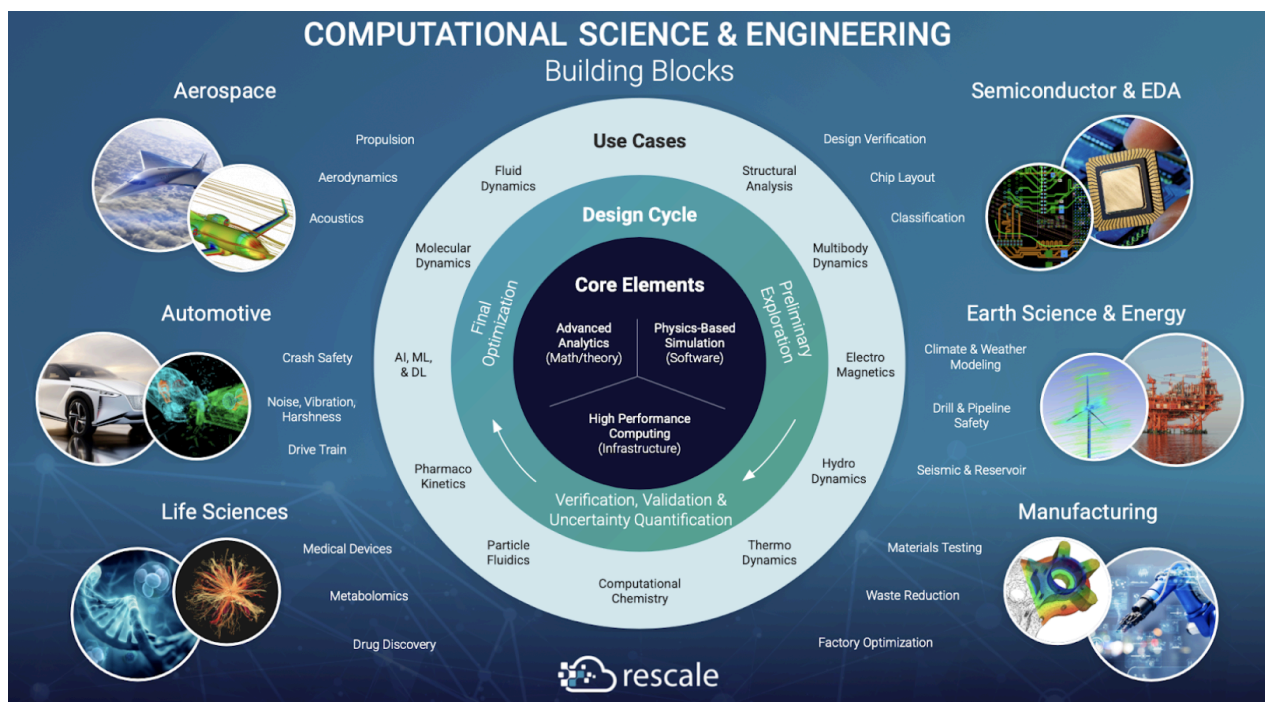
CAE workflows typically involve multidisciplinary simulations, including thermal, structural, and aerodynamic analyses, which are crucial for the development of high-performance products. However, managing these workflows can be incredibly complex due to disconnected tools, fragmented datasets, and manual processes. These challenges hinder the efficiency and accuracy of the simulation process, leading to delays, higher costs, and potentially suboptimal designs. In many organizations, CAE tools for thermal, structural, and aerodynamic analysis are often not integrated with one another, leading to isolated systems that don't communicate efficiently. For example, a thermal simulation tool may not seamlessly share data with a structural analysis tool, making it difficult to account for how heat distribution affects the structural integrity of a component. Similarly, aerodynamic simulations often require data from structural models to understand the impact of airflow on materials and designs. When tools do not communicate properly, engineers are forced to manually transfer data between systems, which increases the potential for human error and data inconsistencies. This lack of integration significantly hampers workflow efficiency, making it difficult to iterate designs rapidly or ensure that all aspects of the product work together harmoniously.

The issue of fragmented datasets also arises in the absence of a centralized database or a unified data management system. In the absence of a shared data repository, different teams working on different aspects of a project often end up with outdated or inconsistent information. This fragmentation can cause misalignment

between the various simulations, ultimately affecting the final product’s performance and quality. Another challenge in managing CAE workflows is the reliance on manual processes to bring together the results of different simulations. Engineers often have to manually adjust models, mesh geometries, and input data across tools, consisting of many manual, repetitive processes that takes time and introduces the risk of errors. As product designs grow more complex, the likelihood of errors and missing context increases, leading to inefficiencies and delays.

Moreover, manual processes can impede the ability to make quick design iterations, which is particularly problematic in industries like automotive and aerospace, where tight development timelines are common. In these industries, the pressure to meet regulatory standards, optimize performance, and reduce time-to-market makes it even more crucial to streamline the CAE workflow.

In industries like automotive and aerospace, where the performance of a product is highly dependent on the integration of multiple engineering disciplines, it is essential for thermal, structural, and aerodynamic simulations to converge seamlessly. For example, in the automotive industry, vehicle designs need to balance thermal efficiency, structural integrity, and aerodynamic performance. A car’s cooling system, for instance, affects its aerodynamics, and any inconsistencies between thermal and structural analyses could result in poor product performance or failure.



Similarly, in the aerospace sector, ensuring that a plane’s structure can withstand aerodynamic forces while managing heat buildup is critical. If each of these analyses operates in a silo, the design may not meet safety or performance standards, potentially putting both the product and its users at risk. As such, efficient workflows that allow for seamless integration of different simulations are not just desirable but necessary.

To address the challenges of disconnected tools and fragmented data, many organizations are now adopting

integrated CAE platforms. These platforms are designed to bring together the various simulation tools into a cohesive workflow, reducing the need for manual data transfer and enabling real-time collaboration across different engineering disciplines. The use of cloud-based solutions and digital twins has also become more prevalent, allowing teams to share and update data instantly, ensuring that everyone is working with the most current information.

Furthermore, automation of routine tasks such as mesh generation, data transfer, and simulation setup is gaining traction. Automated workflows can significantly reduce human error, save time, and improve the consistency and accuracy of results across different analyses. Automation also enhances the ability to perform rapid design iterations, helping engineers identify and address potential issues early in the design process, ultimately leading to higher-quality products and shorter development timelines.

3. Automated Digital Thread: A Foundation for Engineering Agility and Traceability

Cloud-based HPC systems support seamless collaboration by connecting simulations and enabling easy data sharing. This allows distributed teams and value-chain partners to work together efficiently on product R&D. By breaking down analytical silos and providing shared access to computing and data resources, cloud platforms significantly enhance data quality and process efficiency in complex multidisciplinary projects.

Collaborating directly on simulation models using shared resources like compute jobs, interactive workstations, and files removes the need for ad-hoc resource sharing. Streamlined, policy-driven access ensures that all team members have the tools and data they need, boosting productivity and minimizing errors. Shared computing workspaces offer a centralized environment, enabling team members to collaborate effectively, no matter where they are located.

Effectively capturing simulation insights and building a shared understanding is essential for quicker project handoffs and better-informed decisions. Managing collaboration workspaces with customizable fields for consistently captured simulation data and metadata enables easy filtering and searching, making data more discoverable. By sharing context, intent, and results, teams have all the necessary information for decision-making, minimizing data management time and preventing redundant work.

Collaboration on product models happens in real-time with the help of comments and notifications. Team members can request feedback, provide extra context, and document decisions, ensuring key information is captured and shared. Features like pinning important comments, tagging teammates, and linking to other simulation tasks promote effortless knowledge sharing and boost the overall efficiency of the collaboration. Templates for jobs and workstations help speed up repetitive workflows and best practices, boosting R&D productivity.

Application specific policy-driven automations across workspaces and projects simplify user onboarding and enhance overall process efficiency. By standardizing routine tasks and configurations, teams can concentrate on innovation instead of administrative work. Maintaining traceable data and decisions is essential for upholding the integrity of the simulation process. Real-time documentation of decisions through comments and notifications, coupled with links to related simulation jobs, creates a clear audit trail. This traceability enables teams to easily

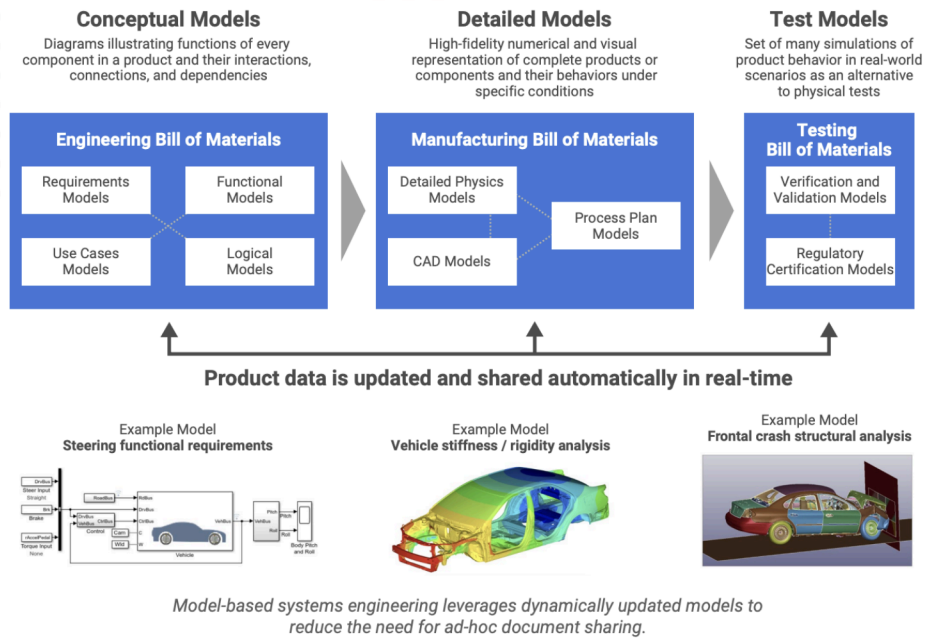
search for and access simulation context, inputs, and findings, ensuring that all decisions are thoroughly documented and readily available.

Enhancing the quality of simulation insights through policy-driven governance, analytics, and simulation templates ensures data consistency and reliability. By programmatically synchronizing trusted data sources like PLM and SPDM systems, accurate and up-to-date information is maintained. Additionally, metadata enrichment helps bridge gaps in context, ensuring data remains relevant and actionable across projects. These platforms often face challenges such as outdated information, inconsistent data values, or a lack of project-specific context, which governance frameworks help mitigate. This governance can be applied at various levels—project, team, or organization—offering administrators the flexibility to enforce only the requirements necessary for engineers to work efficiently.

Shared access and shared context are critical in helping teams establish a holistic view of their product development projects. As relevant product information becomes available, engineers and their stakeholders can aggregate insights from a variety of associated design explorations, experiments, and optimizations to facilitate comparisons of all possible designs. To make this possible, organizations need to ensure data is both accessible and contextualized for all authorized users. Federated storage management and metadata on all simulation outputs can help organize data sets, provide traceability to the models being used, as well as reducing the reliance on anecdotal and ad-hoc document sharing.

The benefit of shared access and shared context for R&D data is that collaboration becomes far easier, faster, and more efficient. Teams can view a single “source of truth” that is accessible and contextualized to help make decisions quickly—regardless of where the data is sitting. As many modern products have an increasingly complex set of requirements and interdisciplinary systems, engineering teams are implementing model-based systems engineering (MBSE) frameworks to streamline data sharing. Using an MBSE approach means engineers can quickly discover the related information they need from the model, as opposed to documents, to perform their design explorations and analyses faster. Handoffs of information between teams are more predictable and key findings remain intact for everyone involved in product development. When the coordination of cross-team collaboration is improved, organizations increase their overall resource efficiency, from reducing duplicated efforts and avoiding unnecessary costs to speeding product development and increasing engineering productivity.

Model-Based Systems Engineering Unifies Data and Automates Data Sharing



4. AI-Driven Case Study

AI-driven models, including graph neural networks (GNNs), are transforming CAE within digital threads by accelerating complex simulations and enhancing prediction accuracy. Integrated into digital threads, AI-powered surrogate models streamline workflows by approximating full simulations with high precision while drastically reducing computational time. This seamless data connectivity enables rapid design iterations, improves design space exploration, and shortens development cycles, ultimately driving more efficient and intelligent engineering processes.

A practical use case involves leveraging AI-driven surrogate models to optimize the complex behavior of bipolar plates in proton exchange membrane (PEM) fuel cells. Traditional simulations, requiring 48 hours and hundreds of CPUs based on the runs on Rescale HPC platform, are significantly accelerated by AI methods that predict structural deformation directly from CAD models. This approach showcases how AI can drastically reduce simulation time and computational demands while maintaining high accuracy.

Case Study: Surrogate Model development for Fuel Cell simulation workflow

The three-step multiphysics CAE approach streamlines the design of flow fields in bipolar plates by integrating finite element analysis (FEA), computational fluid dynamics (CFD), and electrochemical modeling. The process begins with FEA simulations, where gas diffusion layer (GDL) compression maps and deformed GDL geometries are generated based on CAD and boundary condition (BC) inputs. In the second step, these deformed GDLs are manipulated within CFD models to simulate electrochemical reactions, ensuring an accurate representation of fuel cell behavior. The final step involves projecting and predicting contact resistance, which provides critical insights into key performance parameters such as resistance levels, raw current density or voltage, pressure drops, and purge sensitivity. This approach enables a comprehensive assessment of fuel cell efficiency, ensuring better design optimization for enhanced durability and performance.

The fuel cell simulation workflow requires extensive preprocessing, particularly in the preparation of CAD models for FEA. This stage is human-resource intensive, involving detailed FEA meshing and simulation, each taking approximately 24 hours to complete. To streamline the process, AI-based rescaling techniques are employed to refine and optimize GDL geometries for CFD analysis. The workflow includes extracting the initial mesh from the FEA input file, obtaining the deformed shape from STL files, and building a point-to-point relationship between the initial and deformed meshes to compute displacement at each grid point. To enhance computational efficiency, mesh decimation reduces the original 1.6 million points to 500K. With a total of 16 simulation samples, including 12 for training, the AI models are trained on an L40 GPU to accelerate predictions. Additionally, topology tools are employed to improve poorly meshed areas, ensuring a consistent and valid mesh for high-fidelity simulations.

5. Business Outcomes

Adapting engineering workflows to project needs involves integrating computational pipelines, automating tasks, and optimizing resource allocation. Automated workflows unify engineering and scientific analyses, enabling seamless execution of modeling and simulation tasks across multiphysics, AI/ML, and other domains. Workload orchestration plays a crucial role in accelerating the adoption of AI-augmented solvers and optimizing computational pipelines for performance and cost efficiency. By standardizing and scaling modular workflows with advanced execution frameworks, organizations can ensure consistency, reliability, and flexibility in complex simulations. Unified computational analysis fosters collaboration by allowing subject matter experts to share insights at every stage, improving design exploration and performance optimization. Additionally, flexible and modular workflow development enables businesses to apply best-fit tools across various R&D projects, ensuring innovation and adaptability.

Data automation is a key driver in modern engineering, capturing insights and creating a digital thread that enhances traceability and reproducibility. Automated computational pipelines extract and tag key performance indicators (KPIs), providing a centralized view of project details, improving data quality, and streamlining processes. Organizations leveraging digital thread concepts have reported significant improvements, such as a 50% reduction in simulation runtimes, leading to faster product development, cost savings, and better resource utilization. Case studies from industries like automotive and aerospace highlight the impact of these strategies, showcasing enhanced collaboration and shorter design iteration cycles. Looking ahead, AI-powered optimizations, such as real-time performance monitoring and predictive analytics, will further transform digital engineering. These advancements will not only accelerate simulations but also reduce reliance on physical prototypes, driving efficiency, sustainability, and innovation in product development.

6. Conclusion

Integrating digital thread and automation technologies empowers engineering organizations to optimize workflows, drive innovation, and maintain a competitive edge. This approach is especially valuable in industries like automotive and aerospace, where strict design and regulatory standards demand precision and efficiency. By transforming simulation data into actionable insights and fostering seamless collaboration, companies can accelerate product development, reduce costs, and enhance design quality. Automated data management ensures traceability and reproducibility, enabling engineers to iterate quickly and make informed decisions. Ultimately, leveraging digital thread technology allows organizations to streamline operations, improve resource

utilization, and bring high-quality products to market faster.

The digital thread workflow begins with defining the project scope and capturing the intent of the work using Rescale Data, ensuring that all relevant information, such as objectives and context, is documented from the start. Within Rescale UX, users can create a New Study, adding descriptions and goals to establish a clear framework before running simulations. Once the study is set up, simulations are executed using Rescale Compute, where users can initiate new jobs and automatically capture key simulation inputs and findings through Rescale Data. Built-in automation and metadata management enable efficient organization, search, and analysis of results, facilitating collaboration and knowledge sharing. With Rescale AI, simulation data is leveraged to train AI models, incorporating training loops to refine predictions. AI simulations can then be run to optimize outcomes based on the initial project intent. Additionally, this structured workflow extends beyond traditional simulations, supporting advanced use cases such as digital twin applications for maintenance, further enhancing efficiency and decision-making.

The Rescale platform enables a comprehensive, automated workflow for teams and companies that rely on high-performance computing (HPC), simulation, and AI-driven optimization. It benefits not only multi-disciplinary teams but also any team that collaborates on complex computational projects, including engineering, R&D, and data science teams working across industries. Companies that require seamless digital thread integration, cross-functional collaboration, and end-to-end automation will find significant value in Rescale's approach.

7. References

1. <https://cottrillresearch.com/various-survey-statistics-workers-spend-too-much-time-searching-for-information/>

Reshaping Simulation Data for an AI Future

John William, Rescale (Netherlands)

Romain Klein, Rescale (France)

Abstract

The convergence of High-Performance Computing (HPC) and Artificial Intelligence (AI) is revolutionizing engineering simulation, marking the dawn of a data-centric era in innovation and optimization. This paper proposes a comprehensive framework for simulation data management tailored to harness the power of AI for engineering applications. The ability to efficiently manage, analyze, and extract insights from high-quality simulation data is pivotal to realizing the transformative potential of this convergence.

Central to the proposed approach is the establishment of a centralized data repository, designed to unify diverse datasets, including experimental data, numerical simulations, and third-party resources. Such a repository serves as the foundation for streamlined data organization and access, enabling engineers to manage the growing complexity of simulation data effectively. Leveraging AI-powered analytics, advanced machine learning and deep learning algorithms can be applied to these datasets to identify patterns, uncover insights, and drive data-driven decision-making. This capability significantly accelerates the design and optimization processes while reducing reliance on trial-and-error methodologies.

Custom machine learning models play a critical role in this framework, offering tailored solutions for predicting performance metrics, optimizing designs, and automating routine engineering tasks. For instance, in the automotive sector, AI-driven simulations can predict vehicle performance, fuel efficiency, and safety parameters with unprecedented accuracy. This not only shortens development cycles but also enhances the overall quality and competitiveness of the final product.

The concept of digital twins is another cornerstone of this approach. By leveraging simulation data to create high-fidelity digital replicas of physical systems, engineers can perform predictive maintenance, optimize system performance, and conduct virtual testing. These digital twins facilitate rapid design iterations, minimize prototyping costs, and accelerate time-to-market. A collaborative environment further enhances this framework, fostering seamless knowledge sharing among engineers and scientists. Such platforms encourage interdisciplinary innovation, amplifying the impact of AI-driven insights across diverse domains.

The integration of cloud-based HPC offers a flexible and cost-effective computing solution, enabling engineers to scale resources dynamically based on project demands. This ensures optimal utilization of computational power while maintaining cost-efficiency. Moreover, cloud-based storage provides scalable solutions for managing large simulation datasets, ensuring secure, centralized access for global teams.

By adopting this data-centric strategy, industries such as automotive, high tech, life sciences, and manufacturing can unlock the full potential of AI to accelerate innovation, improve product performance, and reduce development costs. This framework exemplifies the synergistic possibilities at the intersection of AI, HPC, and engineering simulation, paving the way for a transformative future in digital engineering.

1. Introduction

The rapid advancement of computational technologies has brought about a paradigm shift in how engineers approach simulation and optimization tasks. High-Performance Computing (HPC) has long been the backbone of engineering simulations, enabling complex modeling and analysis at unprecedented scales. However, the integration of Artificial Intelligence (AI) into HPC workflows is now unlocking new possibilities for data-driven innovation.

Simulation data—whether generated from numerical models, physical experiments, or third-party sources—is increasingly recognized as a valuable asset in driving decision-making processes. The ability to manage, analyze, and derive actionable insights from this data lies at the heart of modern engineering practices. This paper presents a comprehensive framework designed to address these challenges by leveraging AI-driven approaches for simulation data management and analysis.

The proposed framework aims to unify diverse datasets into a centralized repository, enabling seamless integration with advanced analytics tools powered by machine learning (ML) and deep learning (DL). By doing so, it facilitates the extraction of actionable insights from complex simulation data, thereby accelerating design optimization and innovation cycles across industries. This paper explores the key components of this framework, its applications in various domains, and its potential to revolutionize engineering practices in an AI-driven future.

2. Related Work

The convergence of HPC and AI has been a topic of significant research interest in recent years. Studies have explored the use of ML algorithms for predictive modeling in engineering simulations [1], while others have focused on optimizing HPC workflows through cloud-based solutions [2]. However, existing approaches often lack a unified framework for managing simulation data across diverse sources and integrating it with AI tools.

Previous work has also highlighted the importance of digital twins in enabling real-time monitoring and optimization of physical systems [3]. While these concepts are promising, their implementation is frequently hindered by challenges related to data management, scalability, and interoperability. This paper addresses these limitations by proposing a comprehensive framework that integrates simulation data management with AI-driven analytics and cloud-based HPC resources.

3. Proposed Framework

The proposed framework consists of five key components: a centralized data repository, AI-powered analytics tools, custom ML models, digital twins, and collaborative platforms. Each component is designed to address specific challenges in simulation data management and analysis while enabling seamless integration with advanced computational techniques.

3.1 Centralized Data Repository

At the core of the framework lies a centralized data repository that unifies diverse datasets from various sources, including numerical simulations, physical experiments, and third-party databases. This repository is designed to handle large-scale datasets efficiently by leveraging cloud-based storage solutions and scalable database architectures. By consolidating data in a single location, engineers can access comprehensive insights into their simulation workflows while reducing the risk of data silos.

At the core of the framework lies a centralized data repository that unifies diverse datasets from various sources, including numerical simulations, physical experiments, and third-party databases. This repository is designed to handle large-scale datasets efficiently by leveraging cloud-based storage solutions and scalable database architectures. By consolidating data in a single location, engineers can access comprehensive insights into their simulation workflows while reducing the risk of data silos.

The centralized repository captures valuable simulation insights through Metadata Management, which delivers a seamless, flexible, and automated framework for engineering knowledge management. This ensures that enterprise simulation data is highly leveraged for collaborative product design, traceable decisions, and policy-driven governance.

The repository provides an engineer-centric experience with turnkey implementation and intuitive collaboration, offers broad compatibility to work with any simulation applications, PLM systems, and data frameworks, and ensures continuous and consistent metadata capture of insights on all simulation activities and users.

It allows for flexible categorization, enabling effortless organization of simulation jobs, workstations, and files with custom tags, simplifying the search, filter, and grouping of resources. Admins can create and deploy simple forms with custom fields for various field types, streamlining data capture, enforcing process compliance, enhancing traceability, and improving data quality.

The repository ensures up-to-date data for multidisciplinary teams by streamlining simulation data capture, providing unified visibility of key performance measures, and enabling programmatic data extraction for improved speed and accuracy.

It facilitates real-time documentation of decisions with comments and notifications, linking simulation jobs with in-line references to other simulation jobs, allowing users to search and find simulation context, inputs, and findings when needed. Finally, it enhances the quality of simulation insights through policy-driven governance, analytics, and simulation templates, programmatically synchronizing with the source of truth by connecting to PLM and SPDM systems, ensuring data consistency and quality.

3.2 AI-Powered Analytics

The framework incorporates advanced machine learning (ML) and deep learning (DL) algorithms for analyzing

simulation data, enabling automated pattern recognition, anomaly detection, and predictive modeling, thereby reducing reliance on manual analysis techniques.

For example, clustering algorithms can identify trends in large datasets, while classification models can predict the likelihood of specific outcomes based on historical data. The analytics component allows users to centralize, explore, and analyze large volumes of simulation data, discovering insights through the in-app Data Explorer, accessing data programmatically via the Data Access API, or integrating with business intelligence tools to compare multiple simulations.

This comprehensive approach to AI-powered analytics enhances the ability to compare and visualize findings from all past analyses, providing a robust foundation for informed decision-making and continuous improvement in simulation workflows.

3.3 Custom ML Models

To address domain-specific challenges, the framework supports the development and deployment of custom machine learning (ML) models tailored to particular engineering applications. These models can be trained on simulation data to optimize design parameters, reduce computational costs, or improve system performance. For instance, in the automotive industry, ML models can predict vehicle performance under various conditions based on simulation inputs.

The framework includes an AI physics package that enhances traditional simulation methods with specialized ML algorithms, dramatically accelerating engineering and scientific discovery. It offers a range of tools, including open-source software like TensorFlow, PyTorch, and Caffe, as well as accelerated computing with high-performance GPUs for intensive AI tasks. The framework also provides programmable platform integration and workflow automation through API access, comprehensive data labeling for simulation training data, and pre-built AI training templates for model training and validation.

Additionally, it features a broad AI application and pre-trained model catalog, allowing for the publication of custom-trained models for team use. This flexible and comprehensive approach enables innovation-driven enterprises to harness cutting-edge AI models, tools, and hardware to solve complex problems with new levels of computing speed and efficiency, ultimately supercharging simulation and accelerating breakthroughs in product design, digital twins, drug discovery, and more.

3.4 Digital Twins

Digital twins—virtual replicas of physical systems—are integrated into the framework to enable real-time monitoring and optimization of engineering designs. By leveraging simulation data and AI-powered analytics, digital twins provide insights into system behavior under different scenarios, enabling predictive maintenance and performance improvements. A digital twin is a virtual representation or digital model of a physical object or system that uses real-time data and other sources to mimic its real-world counterpart.

Sensors can be placed on physical assets like turbines, pipelines, or manufacturing equipment to collect data on the status, condition, and performance of the system. This data is then fed into the digital twin model, which uses artificial intelligence, machine learning, and analytics to process the data. The digital twin provides insights on how the physical system is currently operating, detects issues early, and runs simulations to predict future

performance. Digital twins allow companies to virtually test changes to systems or conditions without impacting actual operations. For example, they could simulate how a wind turbine would perform under various wind speeds and weather conditions. The learnings can then optimize real-world assets and operations.

Digital twins are becoming a crucial component of the Internet of Things, allowing remote monitoring, controlling, and optimizing of assets from anywhere. As computing power and AI advance, so will the capabilities of digital twin technology. A digital twin begins with creating a virtual model of a physical object or system, defining both the physical components and the rules or physics that govern how it operates and interacts with the environment. The virtual model is then connected to the actual physical asset through an array of sensors that collect real-time data.

Powerful simulation software runs the digital twin model, leveraging artificial intelligence, machine learning, and analytics to mirror how the physical system is performing in a virtual environment. Operators can then perform “what if” scenarios with the digital twin, adjusting various parameters to predict how changes may impact operations. Over time, the machine learning capabilities of the digital twin model continuously improve by comparing simulations to actual results, allowing the digital representations to become increasingly precise and valuable. Digital twins combine physical world data with advanced analytics to optimize productivity, prevent downtime, and develop new opportunities.

There are various types of digital twins, including product twins, production twins, performance twins, system twins, prototype twins, and composite twins, each serving different business purposes. The concept of digital twins has existed for decades, but recent advancements in IoT sensors, cloud computing, and simulation software have significantly expanded their capabilities and applications. Digital twins are now used for predictive maintenance, virtual commissioning, remote monitoring, product development, training, smart cities, logistics, and healthcare.

The future of digital twins looks promising, with wider adoption, greater intelligence, expanded integration, democratization, customization, lifecycle uses, new business models, and automated actions expected to drive dramatic gains in efficiency, performance, safety, and innovation across many industries.

3.5 Collaborative Platforms

The final component of the framework is a collaborative platform that enables engineers, data scientists, and domain experts to work together on simulation projects. This platform supports real-time collaboration, version control, and knowledge sharing, thereby enhancing productivity and fostering innovation within engineering teams. Collaborative platforms are integral to modern engineering and scientific research, providing a unified environment for teams to work together seamlessly. These platforms facilitate real-time communication, data sharing, and project management, enabling multidisciplinary teams to collaborate effectively regardless of geographical location.

By integrating various tools and applications, collaborative platforms support the entire lifecycle of a project, from initial design and simulation to testing and deployment. They offer features such as version control, which ensures that all team members are working with the most up-to-date information, and role-based access control, which maintains data security and integrity. Additionally, these platforms often include advanced analytics and visualization tools, allowing teams to gain deeper insights from their data and make informed decisions quickly.

The use of collaborative platforms enhances productivity, fosters innovation, and accelerates the development of complex engineering solutions by providing a cohesive and interactive workspace for all stakeholders involved. Flexible and scalable computing platforms further support these collaborative efforts by offering automated, secure high-performance computing for any engineering modeling and simulation application. These platforms redefine high-performance computing with intuitive user interfaces, flexible deployment options, and efficient management across teams, optimizing for both performance and compliance.

By leveraging cloud-based solutions, teams can manage computational and scientific tasks seamlessly, performing simulation workflow enjoying a high-powered desktop or supercomputing cluster experience, available in-browser from anywhere. This comprehensive approach ensures that all aspects of the project are efficiently managed, from software deployment and versioning to hardware orchestration and data storage, ultimately driving innovation and efficiency in engineering and scientific research.

4. Applications in Industry

The proposed framework has wide-ranging applications across industries where simulation plays a critical role in design and optimization processes. For instance:

Automotive Industry

The framework can be used to optimize vehicle performance by analyzing simulation data from various sources, including computational fluid dynamics (CFD) models and physical test results. By integrating these diverse datasets, engineers can gain a comprehensive understanding of vehicle dynamics and aerodynamics, leading to more informed design decisions.

Custom machine learning (ML) models play a crucial role in this process by predicting fuel efficiency under different driving conditions. These models leverage historical data and real-time inputs to provide accurate predictions, enabling engineers to develop more efficient and environmentally friendly vehicle designs. For instance, ML models can simulate how changes in design parameters, such as aerodynamics or engine configurations, impact fuel consumption and overall performance.

In high-stakes environments like motorsports, where milliseconds matter, the framework's AI-powered analytics and simulation capabilities can dramatically reduce research and development cycle times. By automating physics-based AI predictions, engineers can optimize every aspect of vehicle aerodynamics, enhancing performance and efficiency on the track. This approach not only accelerates R&D decision-making but also ensures that vehicles are optimized for peak performance under various conditions.

Overall, the application of this framework in the automotive industry exemplifies how advanced simulation and AI technologies can drive innovation, improve efficiency, and enhance the performance of modern vehicles.

Manufacturing Sector

In the manufacturing sector, the framework supports predictive maintenance of machinery by analyzing sensor data and simulation outputs. By leveraging digital twins, which simulate machine behavior under various operating conditions, manufacturers can proactively detect faults and schedule repairs before they lead to

significant downtime. This approach not only enhances the reliability and efficiency of manufacturing processes but also reduces maintenance costs and extends the lifespan of machinery.

Digital twins enable real-time monitoring and optimization of manufacturing equipment, providing insights into performance and potential issues. By simulating different scenarios, manufacturers can predict how machines will behave under various conditions and make informed decisions about maintenance and operations. This predictive capability is crucial for maintaining high levels of productivity and minimizing unexpected disruptions in the production line.

Furthermore, the integration of high-performance computing (HPC) and AI-based performance optimization allows manufacturers to conduct extensive simulations and analyze complex data sets rapidly. This capability leads to improved efficiency, reduced weight, and enhanced durability of manufacturing components. By harnessing the power of advanced computing and AI, manufacturers can accelerate the development of innovative solutions and optimize their operations for better performance and sustainability.

Overall, the application of this framework in the manufacturing sector exemplifies how advanced simulation, digital twins, and AI technologies can drive innovation, improve efficiency, and enhance the performance of manufacturing processes.

5. Discussion

The integration of AI into HPC workflows presents both opportunities and challenges. While AI-driven analytics tools offer significant potential for accelerating engineering innovation, their implementation requires careful consideration of data privacy, security, and scalability issues. Additionally, the development of custom ML models demands expertise in both machine learning and domain-specific knowledge to ensure accurate predictions and reliable results.

Looking ahead, advancements in cloud computing and edge technologies are expected to further enhance the capabilities of this framework by enabling real-time data processing and analysis at scale. The widespread adoption of digital twins is also poised to transform engineering practices by providing unprecedented insights into system behavior across industries.

6. Conclusion

The proposed framework represents a significant step forward in leveraging AI for simulation data management and analysis in engineering applications. By integrating a centralized data repository, AI-powered analytics tools, custom ML models, digital twins, and collaborative platforms, the framework addresses key challenges in modern engineering practices while unlocking new opportunities for innovation.

As HPC and AI continue to evolve, this framework serves as a foundation for future research and development efforts aimed at optimizing simulation workflows and driving actionable insights across industries. By embracing data-centric approaches, engineers can unlock the full potential of computational technologies to transform the design and optimization processes in an increasingly complex world.

7. References

- [1] Ahsun, Adaan. (2024). Machine Learning Algorithms for Predictive Modeling in Photochemistry and Nanoparticle Engineering. Computational Intelligence and Machine Learning.
- [2] Four Ways to Digitally Transform with HPC in the Cloud
<https://rescale.com/blog/four-ways-to-digitally-transform-with-hpc-in-the-cloud/>
- [3] Mohammad Sadegh Es-haghi, Cosmin Anitescu, Timon Rabczuk, Methods for enabling real-time analysis in digital twins: A literature review, Computers & Structures, Volume 297, 2024, 107342, ISSN 0045-7949,
<https://doi.org/10.1016/j.compstruc.2024.107342>.
(<https://www.sciencedirect.com/science/article/pii/S0045794924000713>)
- [4] The Intelligent R&D Data Platform, <https://rescale.com/platform/data/>
- [5] Solving The Data Challenges of Digital Engineering,
<https://rescale.com/resources/solving-the-data-challenges-of-digital-engineering-ebook/>
- [6] Digital Twin for HPC Engineers, Scientists and Managers, <https://rescale.com/digital-twin/>
- [7] Comprehensive Cloud Platform for Accelerated Modeling & Simulation, <https://rescale.com/platform/>

Leveraging LLMs for Automated Post-Processing of Simulation Output Logs

James Imrie, Rescale (EMEA, United Kingdom)

Abstract

The rise of Large Language Models (LLMs) has opened transformative opportunities across industries and engineering simulation processes. This session delves into the innovative use of LLMs in post-processing simulation software output logs, addressing a critical challenge for end users and enterprise organisations alike: extracting actionable insights efficiently.

Enterprises today generate terabytes of simulation data daily, and there is an increasing need for automation in retrieving simulation results, classifying errors (e.g., mesh inaccuracies, memory limitations, licensing issues), and preparing AI-ready datasets from trusted simulation outputs. LLMs offer a powerful solution, acting as a simulation co-pilot to automate these tasks with precision. This enables enterprises to troubleshoot workflows more efficiently while reducing manual effort.

This paper explores the end-to-end process of deploying local LLMs in simulation workflows, maintaining data privacy and data isolation. It covers key steps, including benchmarking LLM performance, applying continuous integration and deployment (CI/CD) pipelines, and validating customer-specific post-processing use cases. By leveraging these strategies, enterprises can establish a seamless data pipeline that automates error classification and generates insights.

Real-world use cases will illustrate the application of LLM-powered tools on industry-standard HPC simulation software such as (but not limited to) Siemens Star-CCM+, Dassault Systèmes Abaqus, and Ansys Fluent. A live demonstration will further highlight the practical benefits, showcasing how organisations can achieve effective and rapid troubleshooting, reduce reliance on support teams, and increase user autonomy. These improvements not only enhance operational efficiency but also contribute to cost-effective workflows through scalable and structured data strategies.

Beyond the technical implementation, the paper will discuss the broader implications of integrating LLMs into simulation workflows. Enterprises can position themselves for the future by enabling more intelligent data practices while driving innovation and collaboration across multidisciplinary teams.

This session is designed for engineering leaders, IT professionals, and simulation experts looking to enhance their HPC operations through cutting-edge technologies. Attendees will leave with actionable insights on harnessing LLMs to transform simulation data management, reduce inefficiencies, and unlock the full potential of AI in engineering.

1. Introduction

As engineering simulations grow more complex, the need to extract actionable insights from log files becomes critical. This paper presents an innovative approach that leverages locally hosted LLMs to automate and simplify this process on the Rescale HPCaaS Platform

Recent advancements in small to medium-sized LLMs have made it possible to deploy high-performance models on widely available hardware commonly used for simulation software. These models now offer the throughput and accuracy necessary to address the demands of this domain.

The primary challenge addressed is the extraction of actionable insights from complex and often difficult-to-interpret log files. In cloud-based environments like Rescale, where there is a high diversity of hardware, software, and workflows, the manual debugging of errors and warnings often requires significant skill and expertise. By leveraging LLMs, end users can enhance troubleshooting efficiency and improve simulation workflows.

Currently, LLMs are being deployed within Simulation and HPC workflows primarily in support roles, such as knowledge base assistance (e.g., AnsysGPT in 2024), or as intermediaries in human-in-the-loop reinforcement (HITL-RL) processes to integrate human feedback into digital systems. However, the landscape is rapidly evolving, and new opportunities for applying these models continue to emerge.

The objective of this paper is to showcase a targeted use case for addressing a common problem. It outlines the architecture of the solution, the development of its initial implementation, and relevant examples of its outputs during testing, as well as potential future enhancements and applications within the Rescale Platform.

2. Background and Challenges

CAE simulations are becoming increasingly complex. Larger models now span multiple physics domains, software tools, and workflows. This increased complexity directly correlates with larger volumes of generated data. As models scale, so too do the demands on hardware and computational environments, often requiring higher core counts, increased memory, and even multiple nodes to support simulations effectively.

On the Rescale platform, +1,500 applications are supported across a wide array of operating systems, architectures, and hardware configurations. While rigorous testing ensures compatibility across this diverse ecosystem, unexpected errors inevitably arise—often only encountered in live production environments. These errors can lead to delays, increased costs, and a diminished user experience, underscoring the value of solutions that can mitigate these impacts.

The successful operation of CAE software demands significant expertise, particularly in error classification. For many CAE engineers, interpreting log files is a challenging and time-consuming task. What may appear as a

benign warning in one scenario could prove catastrophic in another, especially when combined with other subtle or hidden issues. This ambiguity in error reporting adds another layer of complexity to an already intricate process, highlighting the need for tools that can streamline troubleshooting and improve overall efficiency.

LLMs are particularly well-suited for addressing these challenges, providing automated analysis of log files to generate human-readable summaries and actionable insights. Industry-standard scripting tools, such as Python, enable the integration of LLMs into automated workflows with ease. Libraries like transformers, langchain, and openai facilitate tasks such as log file parsing, error classification, and summarization. These tools not only reduce manual effort but also provide a foundation for building scalable, repeatable workflows.

3. Methodology and Approach

Deployment of a Locally Hosted LLM for Inference

Deploying a locally hosted LLM for inference relies on several core components. The following method represents the current implementation; however, it is not the only way to achieve the desired outcome.

Components Used

- Ollama 0.5.5
- Python 3.10 (via Miniconda)
- Various additional Python modules.

To support more agile testing during development, the following tools were also utilized:

- OpenWebUI
- Docker

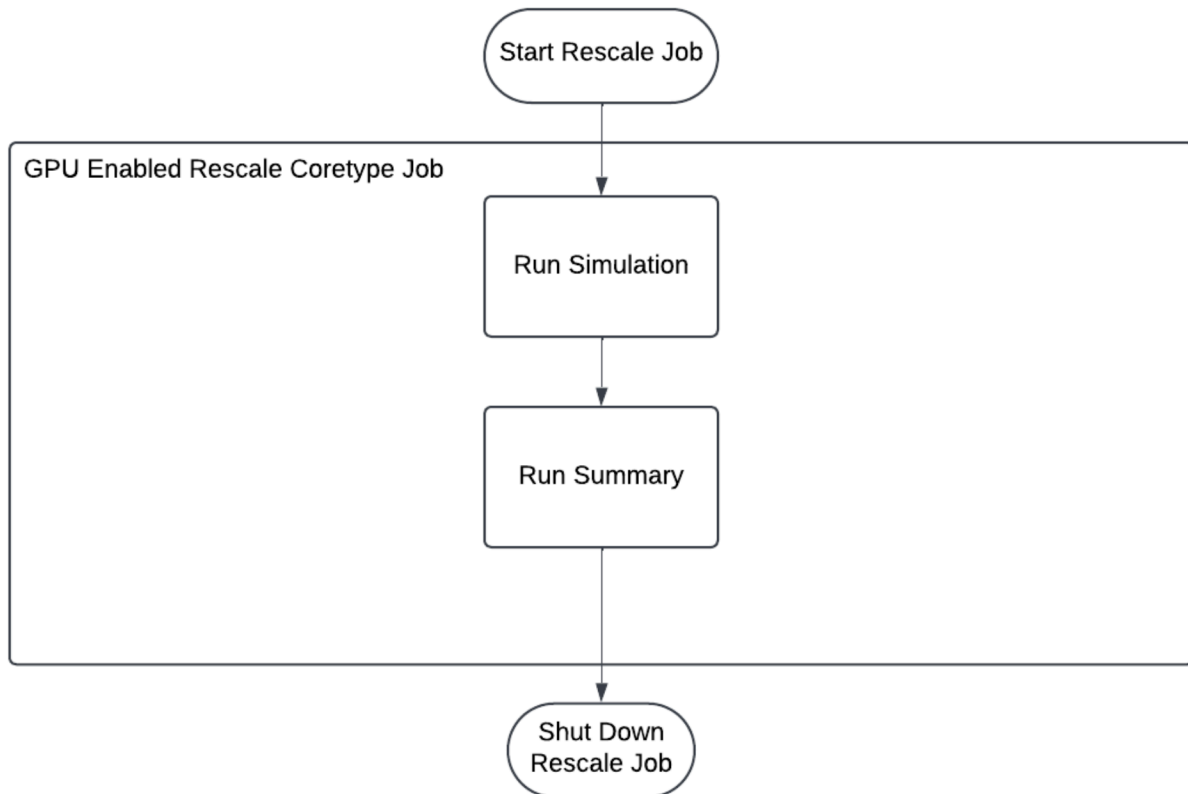
Initial Target Environment

The initial objective for this automation was to integrate the summarization task into the same environment as the simulation process in a sequential process. LLM's typically require a GPU to work with sufficient performance, meaning initially this summarization task can only be performed on jobs where the following is true:

1. GPU Coretype Enabled: NVIDIA A10 or higher.
2. Linux-Based Operating System.
3. Plain text output log files available.

While these criteria narrow the initial scope of application, they were chosen strategically due to the growing adoption of GPU-based simulations. This trend is driven by software vendors increasingly enabling GPU acceleration for their solvers (e.g., Ansys Fluent, Siemens Star-CCM+, and Dassault Systèmes Abaqus). As a result, this initial version retains significant applicability despite its targeted nature.

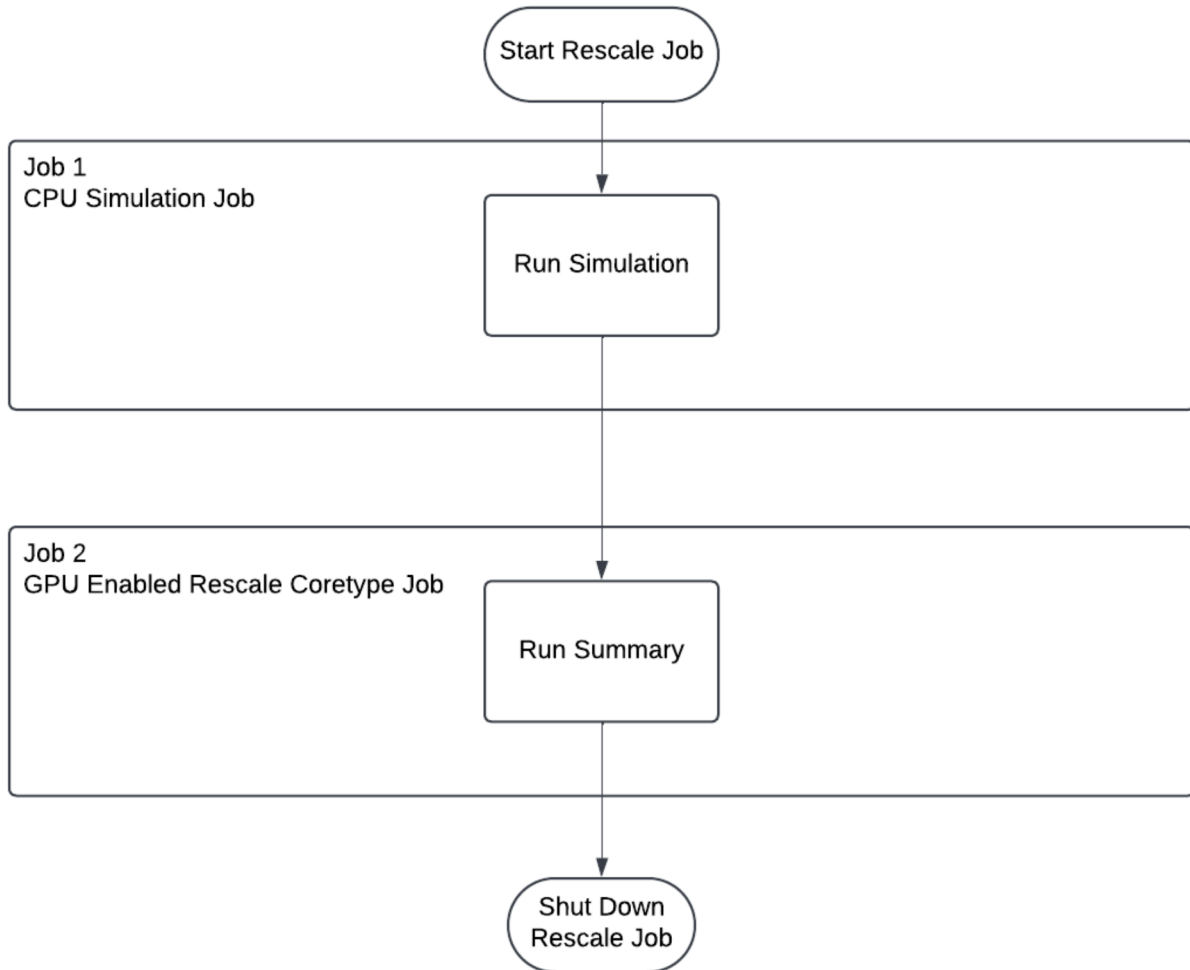
Simplified Process Diagram - GPU



Future Expansion

In future iterations, the automation is expected to extend its compatibility to CPU-based hardware. This will be achieved by creating a dedicated summarization “job” on the Rescale platform, which is launched via an API call from the original job. Decoupling the hardware requirement in this way will increase flexibility and broaden the applicability of the automation across a wider range of workloads, although this will result in a small delay from the simulation job finishing, and the summarization being made available.

Simplified Process Diagram - GPU



Model and Hardware Selection Criteria

The core component enabling this approach is the Large Language Model (LLM) itself. To identify the most suitable model for this task, a research phase was conducted to narrow down the pool of potential candidates. This research provided a shortlist of models for testing, ensuring that only those with the highest likelihood of success were evaluated in the subsequent benchmarking phase.

Initially 3 LLM Models were chosen, these were.

- Mistral-Nemo-Instruct-2407
- Mistral-7b-Instruct
- Phi4

Model Name	Token Count	Context Window	Licence
Mistral-Nemo-Instruct-2407 [4]	12.2B	128k	Apache-2.0
Mistral-7b-Instruct-v0.3 [6]	7B	32k	Apache-2.0
Microsoft-Phi4 [5]	14B	16k	MIT

The models for this application were chosen based on several key factors:

1. Parameter Count

- Parameter count directly impacts the model’s capabilities, task execution speed, and memory requirements.
- Higher parameter counts generally correlate with increased model capabilities. State-of-the-art models are reportedly approaching or exceeding 1 trillion parameters [1]. However, higher parameter counts also result in greater hardware requirements, particularly in terms of memory and GPU VRAM.

2. Context Window

- The context window determines the size of the system prompt, the input file that can be summarized, and overall task performance.
- Models with larger context windows are more capable of handling extensive log files without truncating key information.

3. Quantization

- Quantization affects the model’s performance, memory requirements, and capabilities by reducing the precision of weights. This trade-off is critical for fitting larger models into hardware constraints while maintaining sufficient performance.

4. Public Benchmarks

- Public benchmark results were reviewed to ensure the selected models met the performance and accuracy requirements for log summarization tasks.
-

5. License

- Licensing terms of the models were considered to ensure compliance with commercial usage and redistribution requirements.

Hardware Specification

The hardware chosen for hosting the models was the NVIDIA A10 GPU, known as “Grossular” on the Rescale platform. This selection was driven by its wide availability and the GPU VRAM (24GB GDDR6), which is critical for hosting high-parameter-count models.

The detailed specifications of the “Grossular” Coretype are:

- CPU: 2nd Generation AMD EPYC 7R32.
- Memory (RAM): 8GB per core.
- Storage: 112.50 GiB per core (SSD).
- GPU: NVIDIA A10.
- GPU VRAM: 24GB GDDR6.

This configuration provided the necessary performance and capacity to support both the model’s memory requirements and the inference workload.

Environment Specification

The runtime environment was built with all necessary libraries, drivers, and a compatible Linux-based operating system, internally referred to as “Rescale Linux8.” This operating system is a derivative of RHEL8, customised to work within the Rescale Platform. The Rescale platform fully automates the setup of this environment, ensuring all core dependencies are ready at runtime. Key components included:

1. Ollama

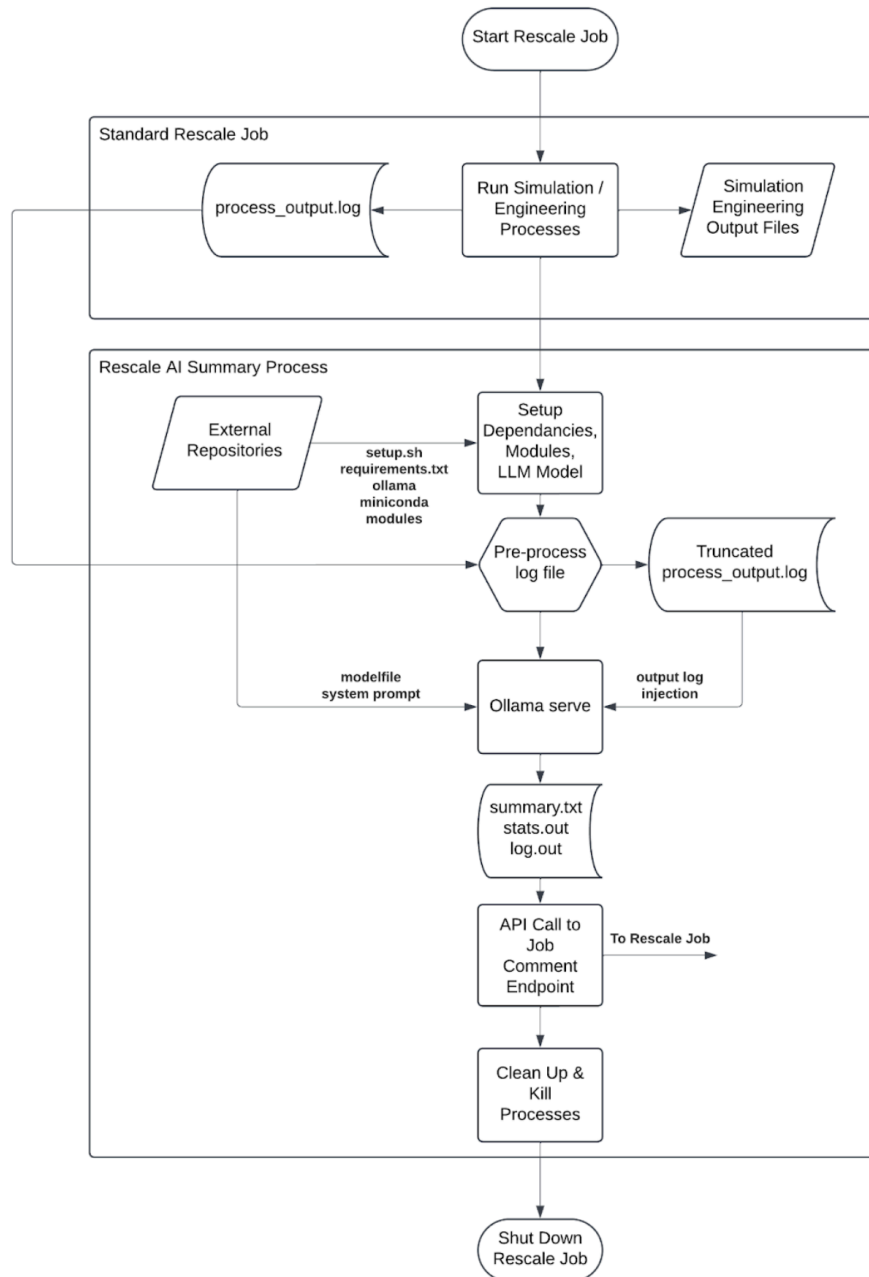
- Purpose:
 - Serves the AI model and simplifies LLM operations.
- Features:
 - Facilitates pulling models from repositories.
 - Serves the models locally.
 - Provides comprehensive logging metrics for performance monitoring.
 - Automatically utilises accelerated computing via GPU libraries and instruction sets.

2. Python

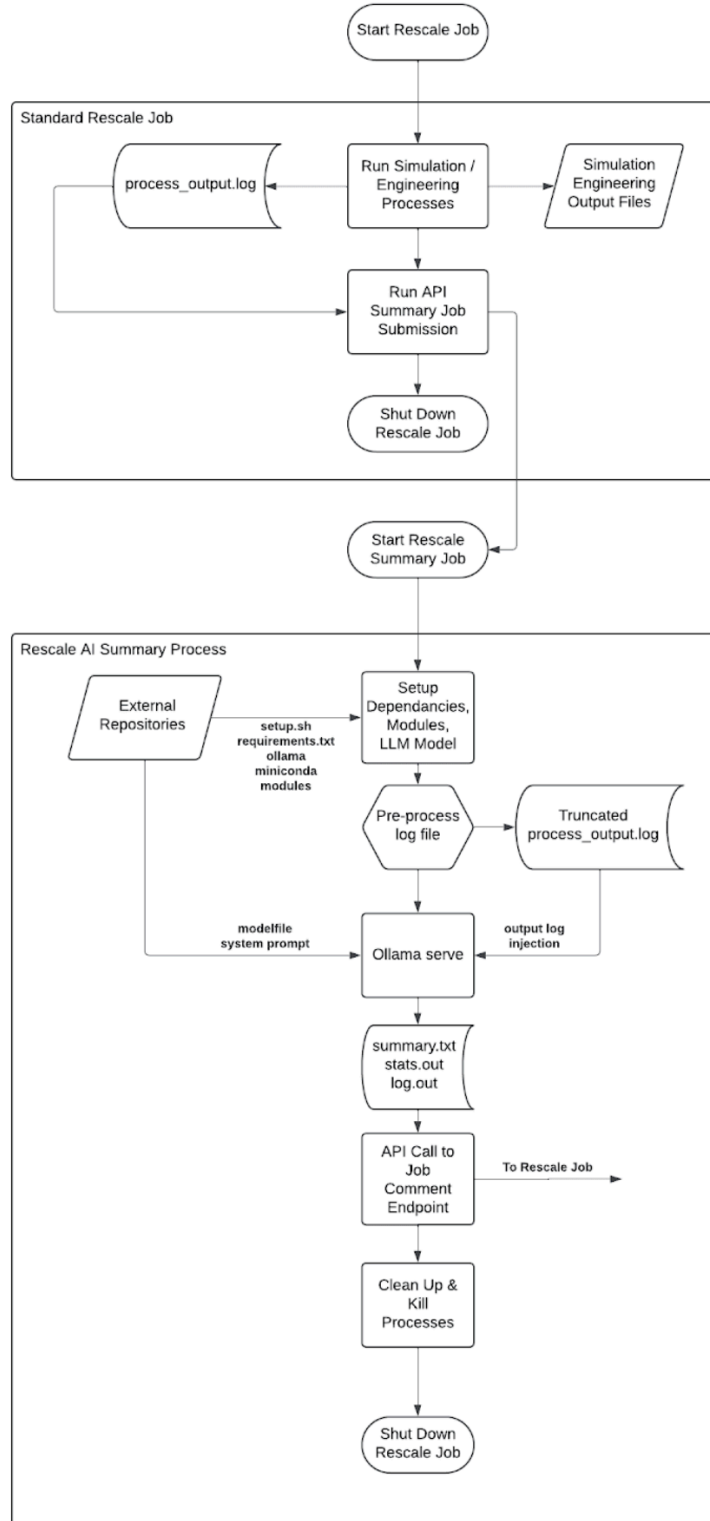
- Reason for Selection:
 - Python was chosen for its compatibility and the availability of a dedicated Ollama module, which abstracts API interactions with the model, eliminating the need for subprocess calls.
- Setup:
 - The Python virtual environment was configured at runtime using Miniconda, with all required modules

installed dynamically during execution.

Process Diagram - Single Environment



Process Diagram - Decoupled Environment



Architecture Commentary

The process has been designed with a modular architecture, ensuring that improvements can be seamlessly deployed without disrupting the overall workflow. This modularity enhances the system's flexibility, maintainability, and adaptability to future developments.

Key Components:

1. Software and Modules

- All required software and Python modules are dynamically pulled from public repositories at runtime.
- Dependencies and versions are explicitly defined in a requirements.txt file to ensure consistency and reproducibility.

2. System Prompt

- The system prompt is sourced from the Rescale Repository at runtime.
- This approach ensures the prompt remains up-to-date, allowing improvements or refinements to be implemented easily.

3. Modelfile

- The model file is also retrieved from the Rescale Repository at runtime, enabling the deployment of newer models as they become available.

4. Drivers, Modules, and Operating System

- The Rescale Platform automates routine updates of drivers, libraries, and the Linux-based operating system to ensure compatibility, stability, and performance.

Security

The process aligns with Rescale's robust security architecture, which adheres to ISO27001 and other industry-standard certifications. This ensures secure handling of system files, authentication, and runtime data.

While the current implementation relies on external data sources for runtime components, these can be eliminated in future iterations by using Rescale's internal tooling. This includes the ability to snapshot required elements or redirect to private repositories for enhanced control and security.

Model Parameters & Testing

The model file in ollama is how model parameters are controlled. [2] The key model parameters that have large influence on the quality of the outputs are:

Quantization

- A 4-bit quantization approach was adopted, striking an optimal balance between performance and memory requirements across the tested models. [3]

Context Length

- A pre-processing step truncates the log file to reduce the input size while retaining critical information for acceptable performance.
- A context length of 8,000 tokens was determined to be sufficient for the task, accommodating both the log file and the system prompt.

Temperature

- A temperature value of 0.2 was consistently used during testing across all models, as recommended in the Mistral Nemo model card.[4]

Seed

- A fixed seed was applied during testing to ensure repeatability of results across all models.

Max Tokens

- The maximum token generation limit was set to 1,000 tokens.
- This ensured responses were succinct yet sufficiently detailed to provide actionable insights to end users.

4. Application & Development Testing

System Prompt

The system prompt plays a critical role in determining the quality and relevance of the output. During development, the prompt refinement process was largely iterative, relying on a trial-and-error approach. By leveraging insights gained from larger commercial models, such as Copilot, the following prompt was crafted and demonstrated consistent reliability across all models tested:

#####

Role Definition:

You are a highly capable and concise process output log summarizer. Your primary task is to analyze the provided log file content, identify any error or warning messages, and accurately determine the process status.

Output Structure:

Present your findings in the following structured format using Markdown, with bold section headers:

1. **Status:**
 - Use a bullet point for the state.
 - Select only from the following: Completed successfully, Completed with warnings, Completed with errors.

2. **Summary:**
 - Use a bullet point for the summary.
 - Provide a brief overview of the process status.

3. **Errors:**
 - Use a bullet-point list to display all error messages.

- Highlight file names, commands, and directories using backticks (e.g., filename.txt, command, /path/to/directory).
4. Warnings:
 - Use a bullet-point list for warnings.
 - Apply the same formatting for file names, commands, or directories as in the “Errors” section.
 5. Actionable Insights:
 - List clear, actionable steps to resolve errors or warnings. If uncertain this should be stated.
 - Use bullet points for each insight.
 6. Potential Search Queries:
 - If unsure about specific issues, provide a list of precise search queries to assist further investigation, with no further explanation
 - Use bullet points for clarity.

Formatting Guidelines:

- Section Headers: Format section headers in bold using **(e.g., **Summary**)**.
- Markdown Syntax: Use appropriate markdown for bullet points (-) and inline code formatting (`) for technical terms, file names, and commands.`
- Accuracy is Critical: Do not guess or make assumptions. If uncertain, focus on providing web queries.
- Conciseness: Be clear and to the point, but include enough detail to be helpful.
- Neutral Tone: Avoid subjective language; provide factual and actionable information only.

#####

System Prompt Techniques

The system prompt design was critical to achieving high-quality, consistent outputs. Several techniques were employed to optimize its effectiveness:

- Role Definition
 - Explicitly defining the role of the model ensured that it adopted a specialized perspective tailored to the specific task. The prompt framed the model as an assistant proficient in summarizing CAE log files, ensuring relevant and focused outputs.
- Output Structure Specification
 - Predefining the structure of the output helped maintain consistency across responses. This consistency made the outputs easier to use in downstream processes and simpler to parse and interpret programmatically.
- Markdown Formatting
 - The Rescale platform supports Markdown rendering in its Front-End GUI. Incorporating Markdown formatting in the system prompt enhanced readability and usability, particularly for technical terms, file directories, code blocks, and inline references. This feature improved the end-user experience significantly.
- Multiple Choice Options
 - By specifying a limited set of predefined outputs for certain sections, ambiguity in responses was

reduced. This approach minimized cognitive load on users interpreting statuses and ensured actionable, structured outputs.

- These predefined options also facilitated tagging operations in subsequent processes, enabling better integration into workflows.

Testing and Refinement

The prompt refinement process was largely iterative and relied on subjective evaluations of whether the responses were useful and accurate. To streamline testing, open-webui was utilized. This tool provided an intuitive interface for dynamically modifying parameters and experimenting with model behaviors on-the-fly. The flexibility of open-webui allowed for rapid testing of adjustments, significantly accelerating the development process.

Test Inputs

To evaluate the performance of the system, a variety of process_output.log files were sourced from both successful and erroneous jobs spanning numerous software packages and workflows available internally at Rescale. With the support of the Rescale Support team, detailed metrics on job failures, software packages, and corresponding rectification steps were compiled. This collaboration enabled the creation of a comprehensive dataset for testing the models.

- Star CCM+
- Ansys Fluent
- Ansys Mechanical
- Dassault Systems Abaqus
- Altair Optistruct
- LS-DYNA

Dassault Systemes Abaqus

process_output_truncated.log from failed Abaqus Job

```
Launching # For single precision analysis, remove double=both
abaqus job=_Full-Displ_AL_mini cpus=$RESCALE_CORES_PER_SLOT
mp_mode=mpi double=both interactive, Working dir:
/enc/udeprod_Deopeb/work/shared. Process output follows:
Analysis initiated from SIMULIA established products
Abaqus JOB _Full-Displ_AL_mini
Abaqus 2024
USING MEMORY DOUG-LEA ALLOCATOR
Abaqus Error: Unable to create temporary directory:
/enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654.
Please check if the name is valid and the permissions for the
directory are correct.
Abaqus/Analysis exited with errors
Exited with code 1
```

Example 1 Analysis: What went wrong?

The Abaqus job was launched with an explicit directory string for the temporary directory that did not exist. This resulted in the job failing with an exit code 1. The error message indicated that Abaqus was unable to create the required temporary directory due to either an invalid path or insufficient permissions.

Example 1 Resolution: How would this be solved?

Remove the explicit path from the Abaqus command, allowing the Abaqus process to create the temporary directory in the valid default location.

Example 1 - Mistral-Nemo-Instruct-2407 [6]

#####

Status:

- Completed with errors.

Summary:

- The Abaqus job _Full-Displ_AL_mini failed to complete due to an error in creating a temporary directory.

Errors:

- Unable to create temporary directory: /enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654. Please check if the name is valid and the permissions for the directory are correct. [process_output_abaqus_error copy.log]

Actionable Insights:

- Verify that the specified temporary directory /enc/udeprod_Deopeb/work/tmp/ exists.
- Check the permissions of the directory to ensure Abaqus has write access.
- Confirm that the generated temporary directory name does not exceed any file name length restrictions.

Potential Search Queries:

- "Abaqus error creating temporary directory"
- "Abaqus job failed due to invalid directory name"

#####

Example 2 - Microsoft-Phi4 [5]

#####

Status:

- Completed with errors

Summary:

- The process encountered an error during execution, preventing successful completion.

Errors:

- Abaqus Error: Unable to create temporary directory: /enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654. Please check if the name is valid and the permissions for the directory are correct.
- Abaqus/Analysis exited with errors
- Exited with code 1

Warnings:

- No warnings were identified in the log.

Actionable Insights:

- Verify that the temporary directory /enc/udeprod_Deopeb/work/tmp/ exists and has appropriate permissions set to allow writing.
- Ensure there are no naming conflicts or invalid characters in the directory path udeprod_Deopeb_Full-Displ_AL_mini_18654.
- Check user permissions for creating directories within the specified working directory.

Potential Search Queries:

- "Abaqus error unable to create temporary directory"
- "Fix Abaqus temporary directory creation issue"
- "Check directory permissions Linux"

#####

Example 3 - Mistral-7b-Instruct-v0.3 [7]

#####

Status:

- Completed with errors

Summary:

- The process encountered an error during the execution of Abaqus job _Full-Displ_AL_mini. The error was related to creating a temporary directory, specifically /enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654.

Errors:

- Abaqus Error: Unable to create temporary directory: /enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654
- Abaqus/Analysis exited with errors
Exited with code 1

Actionable Insights:

- Check the name and permissions of the directory /enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654. Ensure it is valid and has the necessary permissions for Abaqus to create a temporary file within it.
- If the issue persists, consider creating the directory manually with appropriate permissions before running the Abaqus job again.

Potential Search Queries:

- "Abaqus error: unable to create temporary directory"
- "/enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654 permissions"

#####

Key Observations

- Both Mistral variants did not include the "Warning" section, as there were no warnings, whereas phi4 indicated that no warnings were present.

- Mistral7B had issues conforming to the specification of the formatting and included numbers, as well as bullet points.
- All models provided search queries, but the suggestion from Mistral7B was unlikely to yield any positive results due to the overly specific nature of the query.
- All models provided actionable insights, with good guidance.
- All models correctly identified the error and provided some suggestions to help the user.
- All models correctly provided a status from the options provided in the prompt.
- All models provided a summary, but subjectively Phi4 provided a better, more succinct summary that did not duplicate information found elsewhere in the response.

5. Result

The comparison of outputs from the various models tested revealed that the Phi-4 model consistently delivered the most coherent and useful summaries. It followed the requirements defined in the system prompt with a high degree of accuracy, producing outputs that were both well-structured and highly informative.


The example presented in this paper represents just one of many test cases; however, across the majority of the log files analyzed, Phi-4 demonstrated superior performance. Its ability to maintain context, classify errors accurately, and present insights in a clear, actionable format made it the most reliable model for this application.


6. Deployment


Deploying this technique on the Rescale platform is a straightforward process for users. Through the Automations interface, users can simply select the AI Summary Automation during job setup, allowing the summarization process to run automatically.

To streamline adoption, this setup can be templated, ensuring that the automation runs for every job created from the template. This reduces manual configuration and guarantees consistency across multiple job executions. The Comments section (shown in the image below) of each job, is part of Rescale's collaboration feature set, and serves as the designated location for storing these summaries. The summary is automatically posted after the job completes via an API call, leveraging Markdown for rich formatting. Since the comment is intrinsically linked to the job, users can seamlessly access the summary without losing context, improving workflow efficiency and traceability.

Comments

 **Abaqus Example Job**

 **Rozenite**
120 cores

 **Ollama_Test-0.5.4**

Wall time: 2 hrs

Created by: jimrie@rescale.com
Created on: 13/01/2025, 17:58:24

James Imrie Job Owner Me a minute ago

Status:

- Completed with errors

Summary:

- The process encountered an error during execution, preventing successful completion.

Errors:

- Abaqus Error: Unable to create temporary directory: `/enc/udeprod_Deopeb/work/tmp/udeprod_Deopeb_Full-Displ_AL_mini_18654` . Please check if the name is valid and the permissions for the directory are correct.
- Abaqus/Analysis exited with errors
- Exited with code 1

Warnings:

- No warnings were identified in the log.

Actionable Insights:

- Verify that the temporary directory `/enc/udeprod_Deopeb/work/tmp/` exists and has appropriate permissions set to allow writing.
- Ensure there are no naming conflicts or invalid characters in the directory path `udeprod_Deopeb_Full-Displ_AL_mini_18654` .
- Check user permissions for creating directories within the specified working directory.

Potential Search Queries:

- "Abaqus error unable to create temporary directory"
- "Fix Abaqus temporary directory creation issue"
- "Check directory permissions Linux"

Reply Edit Delete 🔖

7. Future Development & Conclusions

The AI Summary methodology has been designed to be modular and highly flexible, allowing it to extend beyond its current scope. Operating within the Rescale Automations Framework, it can be adapted to process any text-based input in an automated manner. Given that text-based log files, metadata, and structured databases are generated throughout various stages of a Rescale job lifecycle, the potential applications of this approach are vast.

Decoupling Model Processing from Simulation Workflows

A key focus for future development is the decoupling of LLM processing from the engineering and simulation workflows. By separating these components, the AI Summary process could be executed as an independent,

scalable service, rather than being tightly coupled to individual jobs. Rescale’s REST API provides a solid foundation for this, offering extensive automation capabilities that facilitate seamless integration with external systems, data sources, and custom workflows.

Harnessing Continuous Model Advancements

The effectiveness of this methodology is inherently tied to the continuous evolution of locally hosted open-source models. Even as of this writing, new models are being released with improved capabilities in reasoning, contextual understanding, and summarization quality. Integrating these advancements into the existing framework can further enhance the model’s performance and output quality.

Exploring Knowledge Augmentation via Web Search

While small to medium-sized models offer impressive performance for this task, they inherently lack the broad knowledge base available to larger foundation models. A promising avenue for future research is the integration of web search augmentation. The current summarization process already generates contextually relevant search queries; the next logical step is to automate these queries, retrieve potentially useful information, and synthesize the results into the summary when applicable.

This augmentation could provide more contextually rich insights without the need to increase model size or hardware requirements. However, this approach introduces challenges, particularly in maintaining the quality and relevance of the summaries while incorporating largely unstructured and potentially irrelevant data. Developing effective filtering mechanisms and relevance-scoring techniques will be critical to ensuring the reliability and usefulness of the augmented outputs.

As these models evolve, they can be seamlessly incorporated into the existing Rescale framework, ensuring continuous performance improvements without major architectural overhauls.

By leveraging automation, scalability, and the rapid innovation in AI models, this methodology has the potential to become an indispensable tool for engineers and researchers seeking to extract meaningful insights from complex workflows more efficiently.

8. References

[1]J. Howarth, “Number of Parameters in GPT-4 (Latest Data),” Exploding Topics, Aug. 06, 2024. <https://explodingtopics.com/blog/gpt-parameters> (accessed Feb. 06, 2025).

[2]“Understanding Modelfile in Ollama,” Pace 'n Think, Apr. 02, 2024. <https://www.pacenthink.io/post/understanding-modelfile-in-ollama/> (accessed Feb. 06, 2025).

[3]xinyuwei, “Phi-4 quantization and inference speedup,” TECHCOMMUNITY.MICROSOFT.COM, Dec. 23, 2024. <https://techcommunity.microsoft.com/blog/machinelearningblog/phi-4-quantization-and-inference-speedup/4360047> (accessed Feb. 06, 2025).

[4]“mistralai/Mistral-Nemo-Instruct-2407 · Hugging Face,” Huggingface.co, 2024. <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407> (accessed Jan. 31, 2025)

[5]“microsoft/phi-4 · Hugging Face,” Huggingface.co, Jan. 08, 2025. <https://huggingface.co/microsoft/phi-4> (accessed Jan. 15, 2025)

[6]“mistralai/Mistral-Nemo-Instruct-2407 · Hugging Face,” Huggingface.co, 2024. <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407> (accessed Jan. 31, 2025)

[7]“mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face,” huggingface.co. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3> (accessed Jan. 31, 2025)



The Most Trusted and Secure Platform for Cloud HPC, Data, and AI

Rescale's cloud platform for HPC and AI accelerates product development while improving IT operational efficiency. Rescale delivers full-stack automation and support of the largest selection of HPC resources, data management tools, and AI capabilities in one integrated platform.



Optimize your HPC

Accelerate full-stack modeling and simulation with Rescale Compute™



Harness your Data

Expedite your product decisions with Rescale Data™



Unlock AI Insights

Discover new products at unprecedented speed with Rescale AITM

Learn more about Rescale

www.rescale.com

[contact us](#)