MAchinE Learning for Scalable meTeoROlogy and cliMate

MAELSTROM

Architecture Blueprint and Solution Design

E4, FZJ

www.maelstrom-eurohpc.eu

D 3.8 Report on solution design and architecture blueprint

Author(s):	FZJ
	Mattia Paladino (E4), Daniele Gregori (E4), Stepan Nassyr (FZJ)
Dissemination Level:	Public
Date:	March 28, 2024
Version:	1.0
Contractual Delivery Date:	31/03/2024
Work Package/ Task:	WP3/ T3.4
Document Owner:	E4
Contributors:	E4, FZJ; All
Status:	Final

MAELSTROM

Machine Learning for Scalable Meteorology and Climate

Research and Innovation Action (RIA)
H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

Project Coordinator: Dr. Peter Dueben (ECMWF)
Project Start Date: 01/04/2021
Project Duration: 36 months

Published by the MAELSTROM Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom
Peter.Dueben@ecmwf.int

The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Unions Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway.

Contents

List of Figures

List of Tables

1 Executive Summary

With the completion of the final benchmarking phase, this deliverable marks the conclusion of Work Package 3 (WP3) within the MAELSTROM project. By synthesizing the findings presented herein, we aim to pinpoint the most effective hardware configurations for the diverse range of applications. Additionally, this analysis will aid in the selection of hardware best suited for the broader needs of the weather and climate (W&C) applications as a collective entity. The insights garnered from this endeavor will serve as the foundation for defining future architectural designs tailored to W&C applications. Central to this document is the utilization of knowledge acquired through an exhaustive benchmarking process. By leveraging these insights, we endeavor to craft a robust and efficient architectural design capable of meeting the rigorous demands of modern W&C activities. Our approach is informed by key findings and outcomes detailed in previous deliverables pertaining to the benchmarking phases. Through this concerted effort, we strive to advance the capabilities and performance standards within the realm of W&C applications.

2 Introduction

2.1 About MAELSTROM

MAELSTROM aims to create Europe's next-generation computer architecture by co-designing custom compute system designs for optimal application performance and energy efficiency, along with a software framework to improve usability and training efficiency for large-scale machine learning applications in weather and climate science.

To achieve this, MAELSTROM will benchmark these applications across various computing systems based on energy consumption, time-to-solution, numerical precision, and solution accuracy. Customised compute systems will be designed that are optimised for application needs in order to enhance Europes high-performance computing portfolio and to pull recent hardware developments towards the unique requirements of weather and climate applications. The MAELSTROM software framework will enable scientists to apply and compare machine learning tools and libraries across a wide range of computer systems with ease. This will be supported by a user interface that links application developers with compute system designers. Also, during the development phase, automated benchmarking and error detection of machine learning solutions will be conducted. These tools will be published as open source.

The MAELSTROM machine learning applications will cover all the key components involved in the workflow of weather and climate predictions. This includes processing of observations, assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of model data and development of forecast products. For each application, benchmark datasets with up to 10 terabytes of data will be available online for training and machine learning tool-development on the fastest supercomputers in the world. The machine learning solutions developed by MAELSTROM will serve as a blueprint for future machine learning applications on supercomputers.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

In the ongoing quest to optimize computational infrastructure for W&C applications, the MAELSTROM project has conducted a series of benchmark analyses to evaluate the performance of various hardware configurations under real-world computational loads. The culmination of this effort is detailed in Deliverable D3.7, marking the third and final phase of our benchmarking endeavors, with previous phases documented in Deliverables D3.4 and D3.6.

Over the course of the project, there has been a notable evolution in both the hardware used for testing and the applications themselves. Initially, our benchmark runs were conducted on NVIDIA V100 and A100 GPUs, representing the cutting-edge technology at the time. However, as the project progressed, we transitioned to testing on newer hardware, including the H100 GPUs, Graphcore IPU and the GH200 superchip, reflecting the rapid advancements in computational technology. Parallel to these hardware upgrades, the applications underwent significant development. Through dedicated efforts from our developers team, each application has been constantly refined and optimized, ensuring that our benchmarks accurately reflect the current state of computational capability and efficiency. For an overview and more detailed information on applications, please consult D1.4.

Given the dynamic nature of both the hardware landscape and application development, this document, Deliverable D3.8, focuses mainly on the data and insights from the most recent benchmarking run reported in D3.7. This approach allows us to base our analysis on the most current and relevant information, ensuring that our conclusions and recommendations are well-founded. Starting from this vantage point, we meticulously analyze the best-performing hardware configurations for each application, aiming to identify the most effective and efficient solutions.

Ultimately, this document will not only highlight the top hardware configurations across various applications but also select the most suited hardware for W&C applications as a whole. It is from this foundation that we will define the architectural blueprint for future W&C applications. This blueprint seeks to leverage the insights gained from our comprehensive benchmarking process to construct a computational infrastructure that is both powerful and efficient, capable of meeting the demanding requirements of modern W&C tasks.

2.2.2 Work performed in this deliverable

The previous benchmarking analysis encompasses a diverse set of experiments to thoroughly evaluate hardware performance under various computational loads.

- Full Training Pipeline: This experiment evaluates the end-to-end performance, from data ingestion and preprocessing through to the training phase, offering a holistic view of application efficiency and hardware utilization.
- Non-IO (Input/Output) Tasks: By focusing on computational tasks without I/O operations, this experiment provides insights into the raw processing power and energy efficiency of the hardware, isolated from the potential bottlenecks of data transfer.
- Inference: Conducted in the context of D3.7 exclusively for an application, the inference experiment evaluates the hardware's ability to execute pre-trained models quickly and efficiently, a crucial aspect for real-time application deployment.

For this deliverable, our analysis focuses on two key metrics: 'Time to Solution', which captures the total execution time of a given application, and 'GPU Energy to Solution', which focuses on quantifying the energy spent by GPUs. These metrics serve as the basis for calculating the 'Action Score', a metric that attempts to harmonise the trade-off between computation speed and energy consumption. This balanced metric is fundamental to evaluating the performance of hardware configurations. Moreover, we performed an analysis to scrutinize and depict the energy-to-solution relative to host power consumption, in conjunction with GPU energy consumption.
Benchmarked devices have different Energy-to-Solution and Time-to-Solution. The overall Energy-to-Solution depends additionally on the power consumption of the Host system and the total run-time of the benchmark. We introduce a new type of plot in which we explore Energy-to-Solution with different accelerators as a function of host power consumption (without the accelerators). This allows for an analysis of whether a tradeoff in favor of Energy-to-Solution against Time-to-Solution still holds for slower, more efficient devices, when integrated in a full system.
For each number of devices that was benchmarked we plot the total Energy-to-Solution for host power draws between 25 and 500 Watt. The best performers show the lowest overall energy. Of special interest are intersection points which indicate where a faster device starts outperforming a slower, more efficient device.

By examining the behaviour of each hardware configuration across the entire training pipeline, Non-IO tasks and inference scenarios, we aim to identify a configuration that not only excels in isolated benchmarks, but also offers the most convincing performance all-round. The insights derived from this comprehensive analysis will guide the formulation of an architectural design specifically tailored to W&C applications. This design is intended to inform the strategic deployment of computational resources, ensuring an optimised blend of performance and energy efficiency that sets a new benchmark for weather and climate infrastructure.

2.2.3 Deviations and counter measures

We noted two minor deviations in our analysis. Initially, we employed a reduced number of hardware configurations compared to those available. However, we still documented a significant number of configurations, particularly those from the latest generation. Secondly, the synthetic benchmarks for the final architecture were not conducted directly by us; rather, they were provided by NVIDIA. Thee results are documented in [2].

3 Benchmark Applications

In the following analysis, we assess hardware performance across the variety of applications defined within MAELSTROM, utilizing specific experiments to discern the configurations that deliver optimal efficiency and overall effectiveness. This section aims to uncover insights into the best-performing hardware for the six MAELSTROM applications AP1-AP6, setting the stage for informed decision-making in system architecture for W&C applications. In order to have a clearer view of what will be described in the next sections, the hardware configurations used for benchmarking are summarised in Table 1. Please note also that the MI250 GPUs are built as Multi Chip Modules (MCM) and because of that they are shown as 8 Graphic Compute Dies (GCDs) with 64 GB memory each.
The reader is referred to D3.7 for more details.

Table 1: Summary of Hardware Configurations

Configuration	CPU	Memory	GPU	Network
JRDC-A100	2x AMD EPYC 7742, 2x 64 cores, 2.25 GHz	512 GB DDR4, 3200 MHz	4x NVIDIA A100 GPU, 4x 40 GB HBM2e	2x InfiniBand HDR (100 Gbit/s)
JRDC-MI250	2x AMD EPYC 7443, 2x 24 cores, SMT-2	512 GiB DDR4-3200 RAM	4x AMD MI250 GPUs, 128 GB each (MCM)	1x Mellanox HDR InfiniBand ConnectX 6 (100 Gbit/s)
JRDC-H100	2x Intel Xeon Platinum 8452Y, 2x 36 cores, SMT-2	512 GiB DDR5-4800 RAM	4x NVIDIA H100 PCIe GPUs, 80 GB each	1x BlueField-2 ConnectX-6 DPU @ EDR (100 Gbit/s)
JRDC-GC200-IPU	2x AMD EPYC 7413, 2x 24 cores, SMT-2	512 GiB DDR4-3200 RAM	4x GC200 IPUs	1x Mellanox EDR InfiniBand ConnectX 5 (100 Gbit/s), 1x Mellanox 100 GigE ConnectX 5
E4-GH200	1x NVIDIA Grace 72-Core	1x LPDDR5X 480GB RAM	1x NVIDIA GH200 GPU	Mellanox CX7 NDR Dual Port interconnection
E4-A2	2x NVIDIA Xeon Gold 6426Y CPU	32x DDR5-4800 32GB	2x NVIDIA A2 Tensor Core GPU	Mellanox CX6 HDR Single Port interconnection

3.1 AP 1

In analyzing AP1, we focus on two key areas: the Full Training Pipeline and Non-IO experiments, providing insights into hardware performance across varied computational scenarios. Since the number of devices for the AMD MI250 represents the number of GCDs, not the number of cards, we have divided this number by 2. This way the same number of individual accelerators is compared.

3.1.1 Full Training Pipeline

The performance of various hardware configurations running the full application pipeline of AP1 are reported in the following figures. Figures 1 and 2 show, respectively, the power consumption of the various GPUs and the total runtime for each configuration varying the number of devices involved. Figure 3 instead shows the scores obtained by each hardware configuration.
The E4-A2 configuration, employed in a single GPU setup, exhibited the lengthiest total execution time and action score among all configurations. In contrast, the E4-GH200 configuration, utilizing a lone Grace Hopper GPU, emerged as the most energy-efficient option. It combined a brief execution time (averaging 458.53s) with minimal power consumption (32.16Wh), resulting in the lowest overall action score among all configurations, totaling 53.092 MJs. Comparing configurations with four GPUs, the JRDC-H100 configuration stood out by achieving a score comparable to that of the Grace Hopper system, with an action score of 57.510 MJs. This highlights the superior performance of the H100's advanced architecture over the previous model, the NVIDIA A100. Then, the JRDC-MI250 configuration showcased a higher action score compared to the NVIDIA GPUs.

3.1.2 Non-IO Experiments

In this set of data from the non-I/O experiments designed to isolate accelerator performance using synthetic data, we observe how different hardware configurations perform under controlled conditions.
The E4-A2 configuration shows a considerable reduction in total runtime (see Fig. 5) compared to its full application pipeline performance, indicating that while still not the most efficient for training tasks, its performance improves significantly when I/O is not a factor. However, the action score (reported in Fig. 6) remains relatively high (92.108 MJs), suggesting that it's not the optimal choice for intensive computational tasks.
The JRDC-H100 configuration demonstrates best performances, both in terms of total runtimes (74.56 s) and action scores (3.27 MJs), highlighting its strong per-

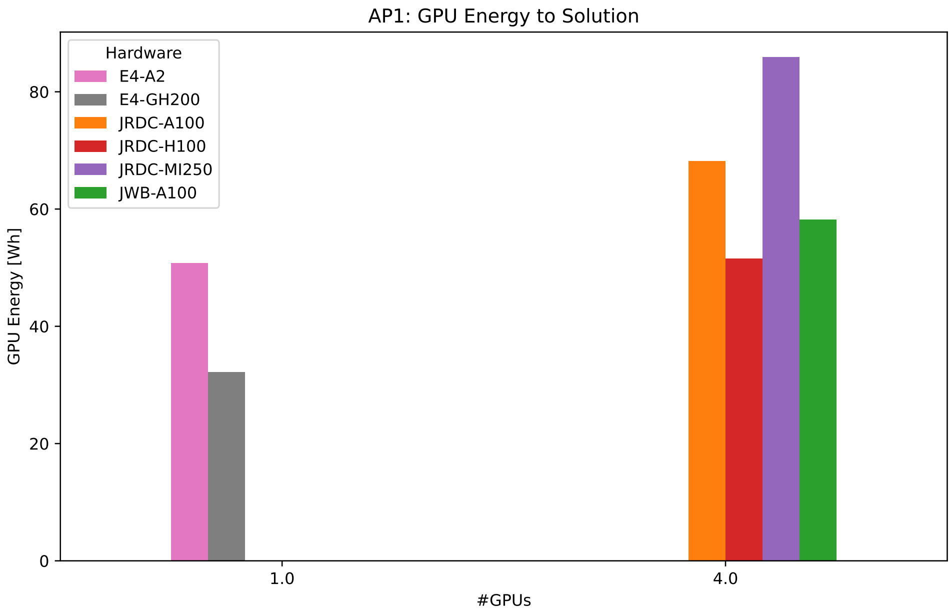
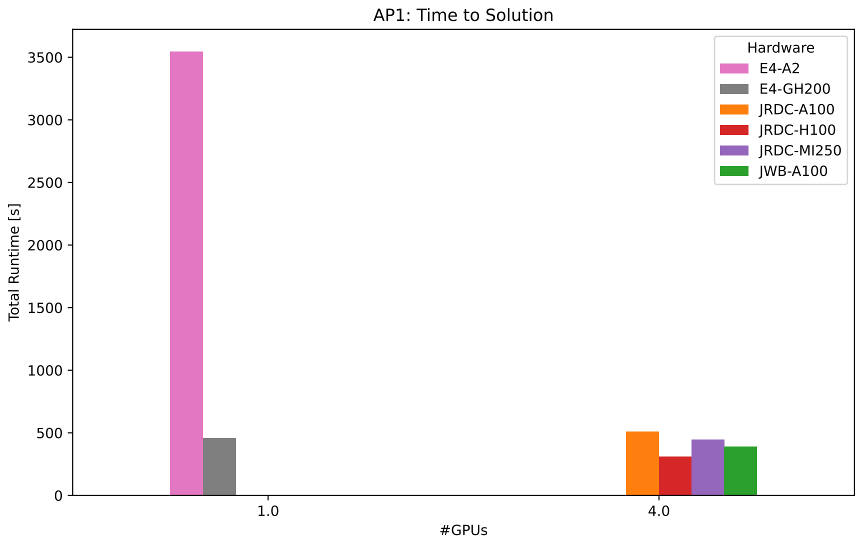
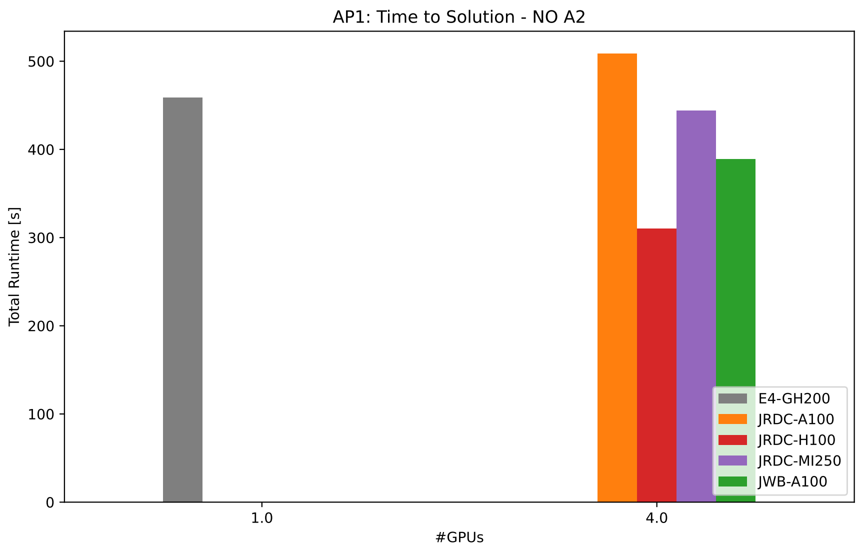


Figure 1: AP1 GPU Energy to Solution: GPU Energy to solution considering the different hardware configurations used. ap1-best-gpu-en

formance in computational tasks even when handling synthetic data. The GH200 showcases good performance, achieving a slightly higher action score (5.667 MJs) compared to the configurations with four devices, while still remaining comparable. This reinforces its position as a highly efficient option for this application.
Notably, the JRDC-GC200-IPU, with its distinct architecture, achieves the lowest energy consumption, that amounts to 7.44 Wh (see Fig 4), and the second best score (3.55 MJs), underlining its potential for specialized tasks where its architecture can be fully leveraged.
The JRDC-A100 and JWB-A100 configurations, both utilizing NVIDIA A100 GPUs, maintain their stance as powerful options for deep learning and scientific computations, though they exhibit a higher action score compared to the GC200-IPU and H100 configurations.
The JRDC-MI250 configuration shows a remarkable ability to handle large effective batch sizes, which could indicate its strength in parallel processing tasks.
Lastly, the JWB-V100 configuration, with its older technology compared to the A100 and H100 GPUs, demonstrates an higher action score with respect the other configurations (except A2), showing improved efficiency and performance offered by newer GPU models.
In summary, this non-I/O experiment data highlights the importance of matching hardware capabilities with application-specific requirements, especially when fo-



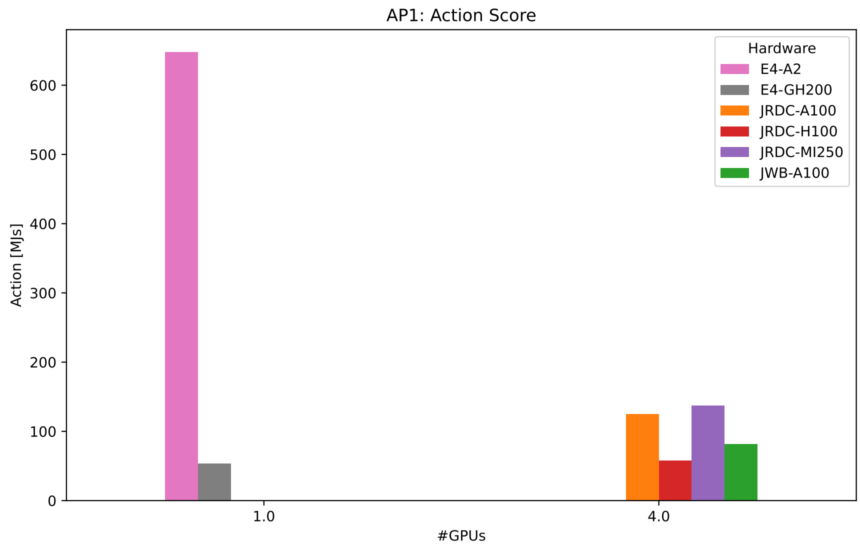
(a) Runtime comparison



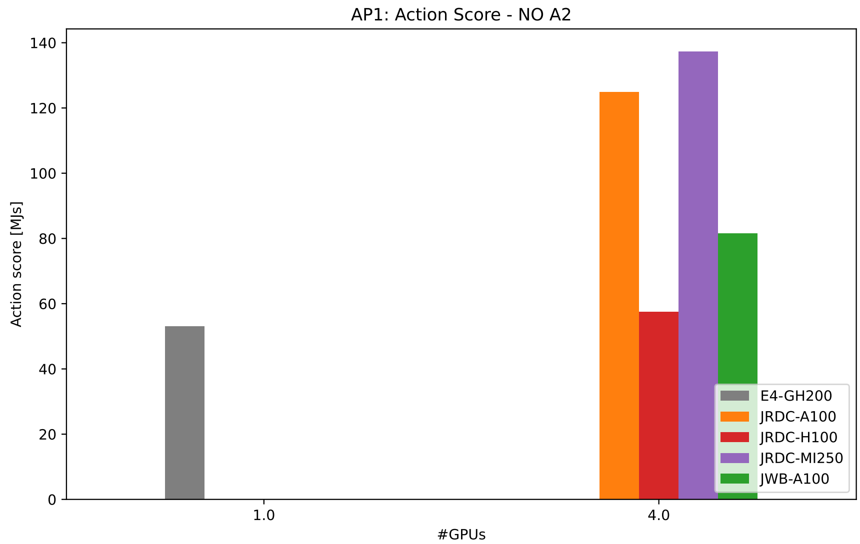
(b) Runtime comparison without A2

Figure 2: AP1 Time to Solution: Time to solution considering the different hardware configurations used. ap1-best-runtime

cusing on accelerator performance. The GH200 and H100 configurations stand out for their efficiency across tasks, while the GC200-IPU and MI250 offer unique advantages for specialized or highly parallelizable workloads. These insights can guide the selection of hardware configurations that best meet the performance and efficiency needs of Weather and Climate applications, balancing the computational



(a) Action comparison



(b) Action comparison without A2

Figure 3: AP1 Action Score: Action Score considering the different hardware configurations used to determine the most performing one. ap1-best-action

demands with energy consumption considerations.

3.1.3 Host Power analysis

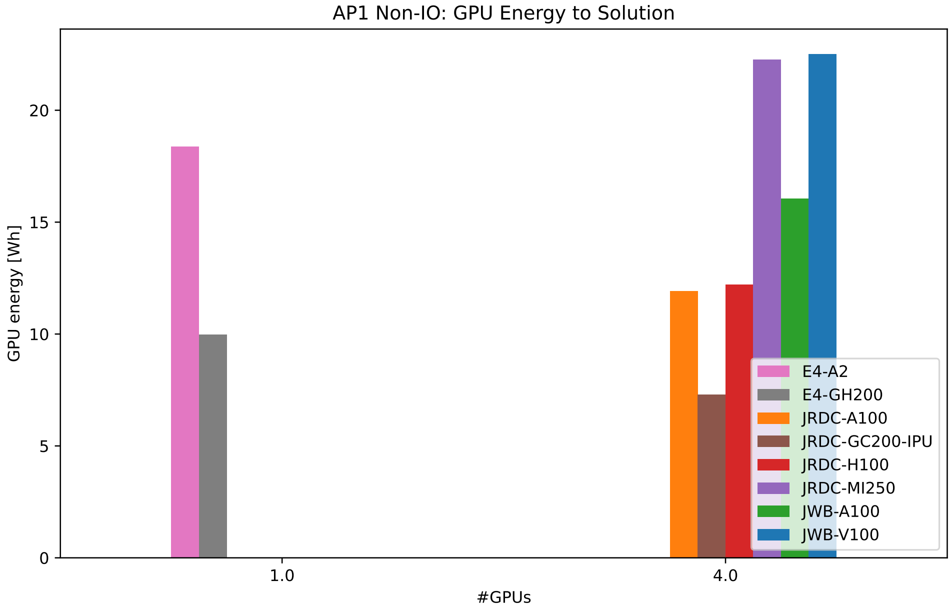


Figure 4: AP1 Non-IO Energy consumption: GPU Energy to solution for Non-IO experiments considering the different hardware configurations used. ap1-nonio-best-gpu-en

3.1.3.1 Full training

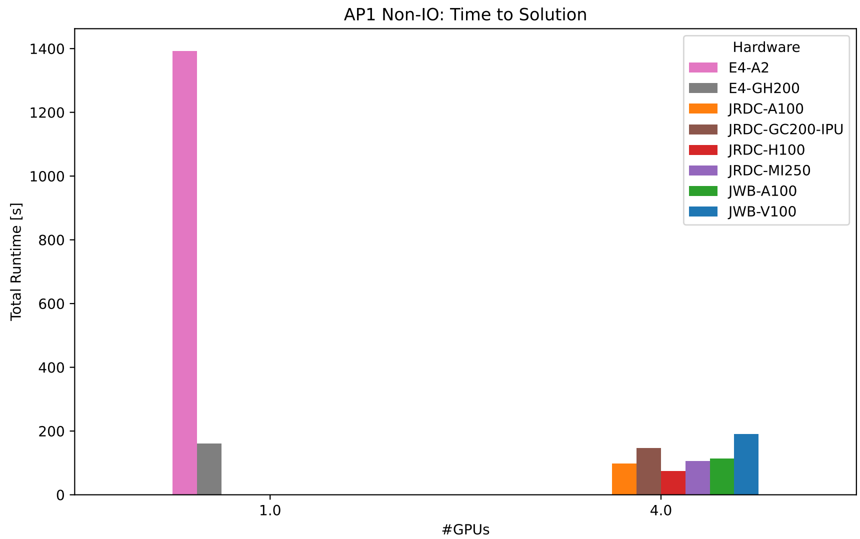
As seen in Fig. 7, for Application 1 we do not see any crossing lines within the investigated host power range. The best performers are the GH200 for 1 GPU and H100 for 4 GPUs.

3.1.3.2 Non-IO Experiments

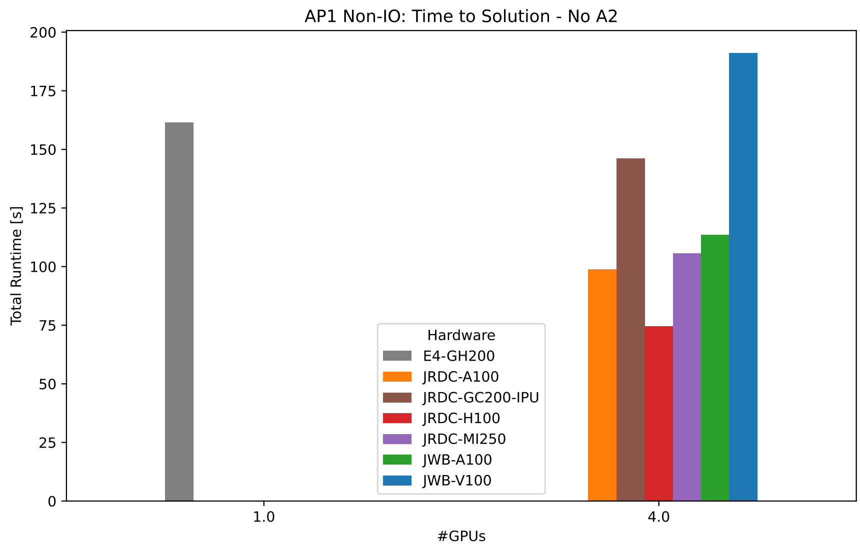
As seen in Fig. 8, we see the Graphcore IPU perform best. However this changes for a host power consumption above 300 Watt, where the H100 starts to perform better with 4 GPUs.

3.1.4 Conclusions

When synthesizing insights from both the training pipeline and Non-IO experiments, the E4-GH200 and JRDC-H100 consistently rank as top performers, showcasing their versatility and efficiency across both full application pipelines and accelerator-focused tasks. The JRDC-GC200-IPU stands out in the non-IO experiments, emphasizing its potential for specific computational tasks where its architecture can be fully exploited.

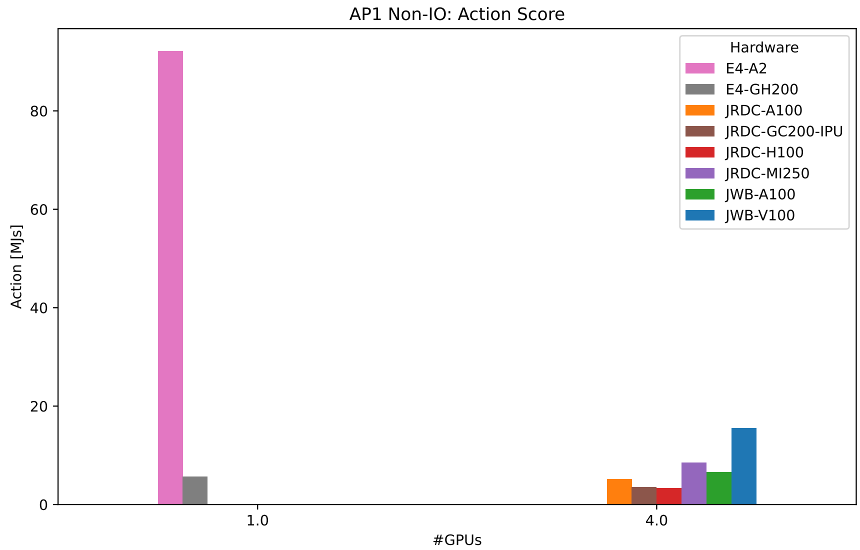


(a) Runtime comparison

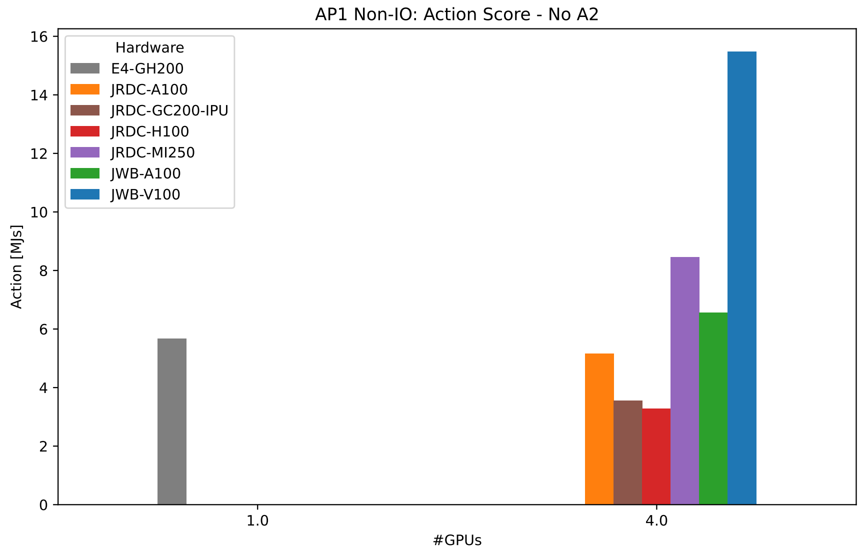


(b) Runtime comparison without A2

Figure 5: AP1 Non-IO Total runtime: Time to solution for Non-IO experiments considering the different hardware configurations used. ap1-best-nonio-runtime



(a) Action comparison



(b) Action comparison without A2

Figure 6: AP1 Non-IO Action Score: Action Score for Non-IO experiments considering the different hardware configurations used to determine the most performing one. ap1-nonio-best-action

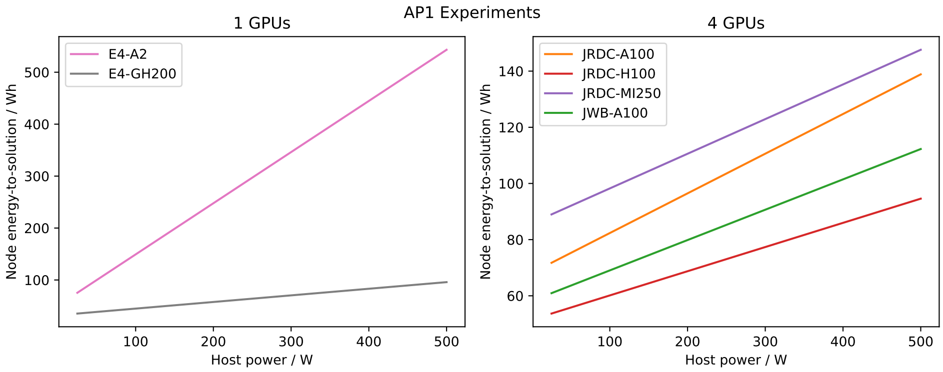


Figure 7: AP1 Energy-vs-Host-Power (Training): Full node energy consumption as a function of host power ap1-tr-etos

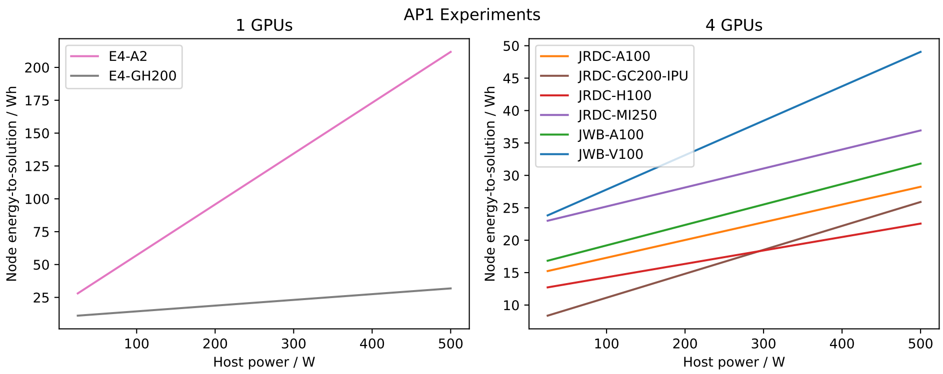


Figure 8: AP1 Energy-vs-Host-Power (Non-IO): Full node energy consumption as a function of host power ap1-nonio-etos

3.2 AP 2

3.2.1 Full Training Pipeline

The data gathered within the scope of AP2 pertains solely to the training pipeline. In figures 9, 10, and 11, we present the GPU energy consumption, total runtime, and action score values, respectively.
The data reveals that increasing the number of GPUs does not significantly decrease runtime, suggesting that AP2 might not effectively utilize multiple GPUs for enhanced parallel processing.
Among the evaluated configurations, the JRDC-H100 with a single GPU stands out for its efficiency, achieving the lowest action score (9.77 MJs) and demonstrating that sophisticated hardware like the H100 can meet the application's demands effectively on its own. The single-GPU JRDC-A100 configuration follows closely in efficiency, further indicating that for AP2, a single GPU is sufficient for optimal performance.
The E4-GH200 also delivers commendable performance in a single-GPU setup, emphasizing its potential for high efficiency in computational tasks. Like the H100 and A100 configurations suggest, additional GPUs do not proportionally improve AP2's runtime, underscoring a scalability issue.

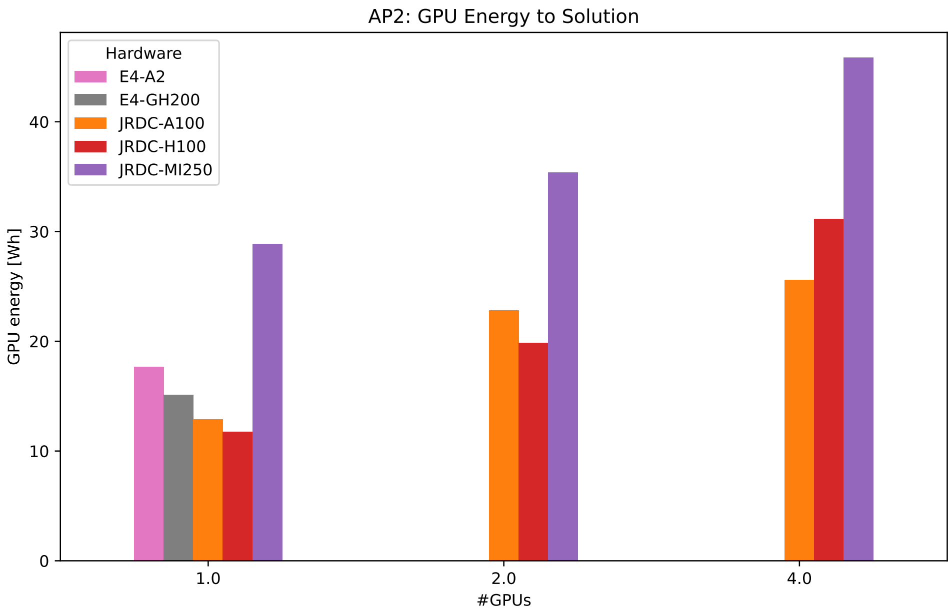
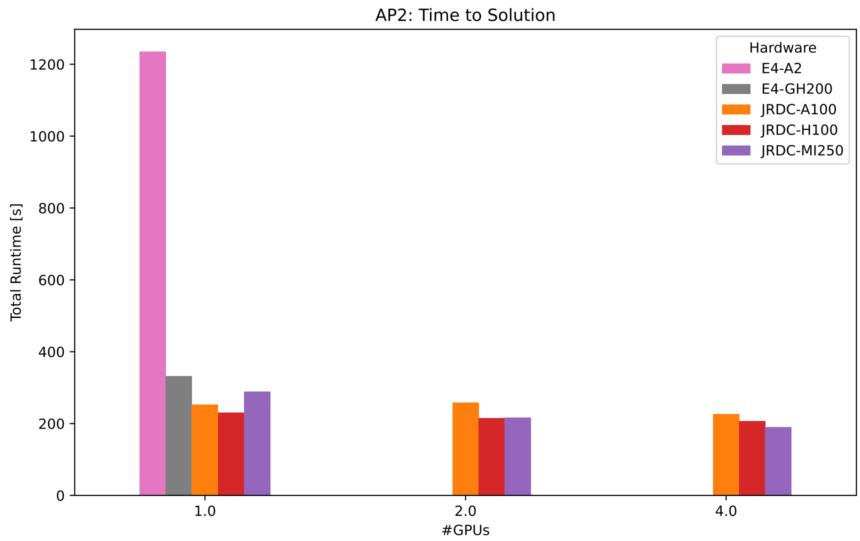
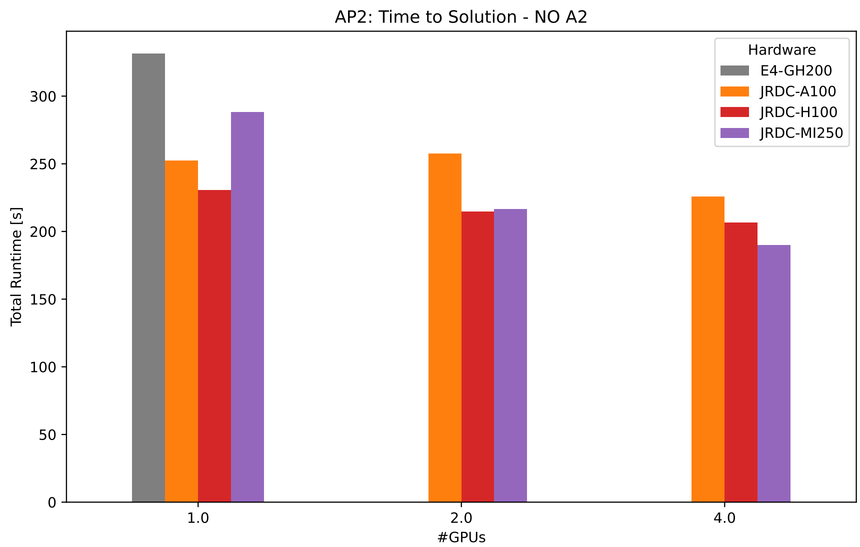


Figure 9: AP2 GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used. ap2-best-gpu-en



(a) Runtime comparison

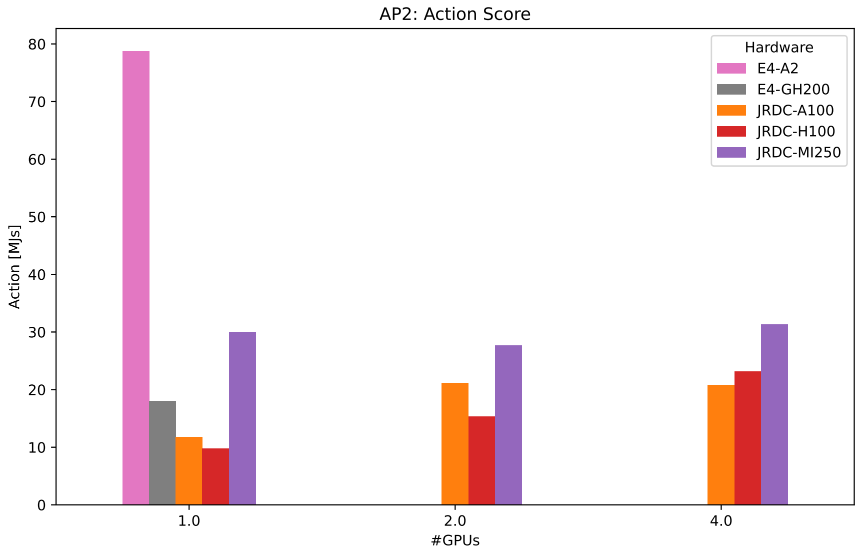


(b) Runtime comparison without A2

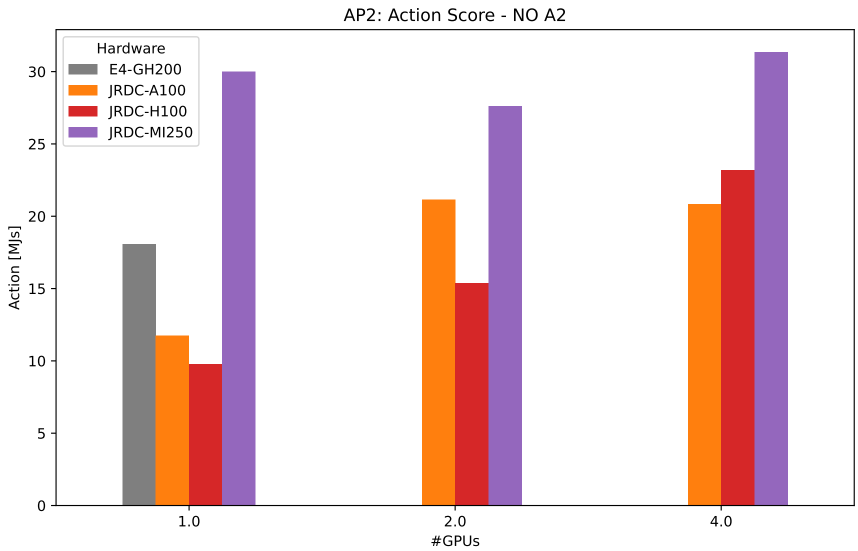
Figure 10: AP2 Time to solution: Time to solution considering the different hardware configurations used. `ap2-best-runtime`

3.2.2 Host Power analysis

As seen in Fig. 12, for Application 2 almost none of the lines cross. The A2 is beat by all other devices except for an extremely low host power consumption of 25 Watt, wher we see the A2 outperform the second highest Energy-to-Solution. An



(a) Action comparison



(b) Action comparison without A2

Figure 11: AP2 Action Score: Action Score considering the different hardware configurations used to determine the most performing one. ap2-best-action

investigation on efficient systems integrating multiple A2s could be beneficial. The best performers are the H100 for 1 and 2 GPUs and the A100 for 4 GPUs.

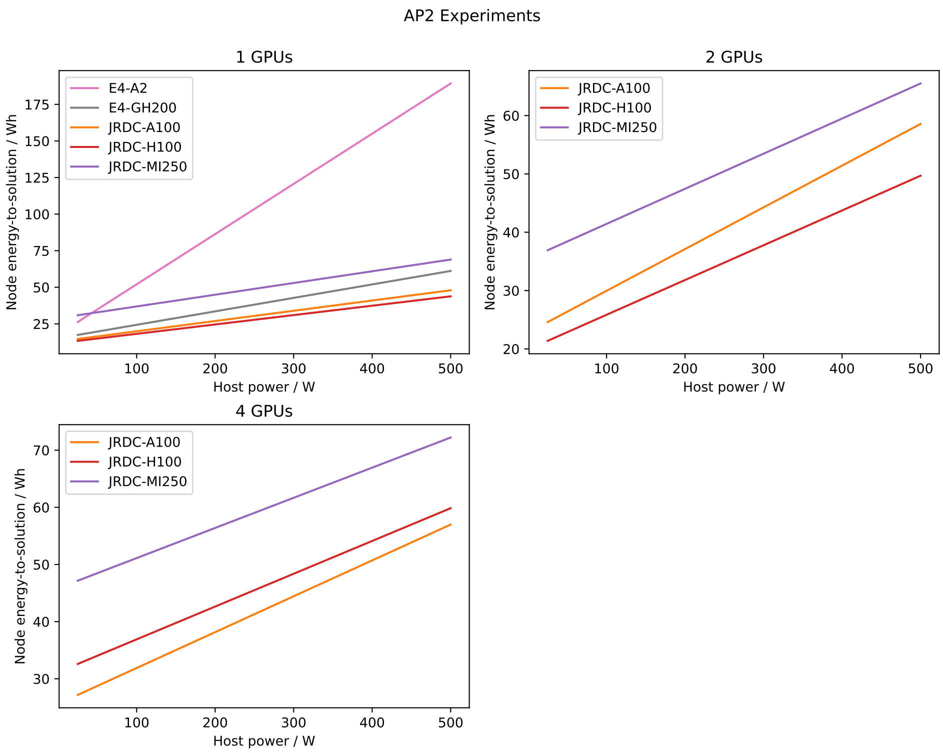


Figure 12: AP2 Energy-vs-Host-Power: Full node energy consumption as a function of host power ap2-tr-etos

3.2.3 Conclusions

The lack of scalability that emerges from the data indicates that AP2's computational workload may not be optimally distributed across multiple GPUs. Consequently, for this kind of application, investing in single-GPU configurations, such as the JRDC-H100 or E4-GH200, may provide the most cost-effective and energy-efficient solution.

3.3 AP 3

In the context of AP3, the analysis performed will include not only the training pipeline, but will also extend to Non-IO and inference experiments. In this way, we aim to obtain a comprehensive understanding of the dynamics of performance in the different phases of the AP3 workflow.

3.3.1 Full Training Pipeline

In the context of AP3's full application pipeline, the JRDC-MI250 with 4 GPUs and E4-A2 stand out as the top-performing configurations. The MI250 setup not only achieves the lowest action score (see Fig. 15) among all tested configurations but also demonstrates the benefits of scalability. Leveraging multiple GPUs significantly enhances the computational speed (see Fig. 14), making it one of the best choice for handling AP3s computational demands.
The E4-A2 configuration showcases high energy efficiency, with an average GPU consumption of 29.56 Wh (see Fig. 13), and good performance, particularly when considering its action score relative to its single GPU counterparts. The E4-A2 offers a viable solution for scenarios with specific energy efficiency and computational requirements. The E4-A2's role is more accurately seen as an efficient single-GPU option, suitable for applications with less intensive computational demands or where energy efficiency is a critical consideration.
In conclusion, the comprehensive analysis positions the JRDC-MI250 with 4 GPUs and the E4-A2 single GPU configurations as the premier systems for AP3s full application pipeline, with Action Scores of 178.85 MJs and 220.20 MJs, respectively.

3.3.2 Non-IO Experiments

We now shift our focus to analyzing the data collected for the Non-IO runs, which include performance data for the GH200 and the IPU Graphcore. In Figure 16, energy consumption data is presented, while Figure 17 displays runtime data. Finally, in Figure 18, the score of each configuration is depicted.
The E4-GH200 stands out as the most efficient option among single GPU configurations, demonstrating its ability to deliver high computational performance while minimizing energy consumption. Transitioning to multi-GPU configurations, the JRDC-H100 with 4 GPUs emerges as the top performer in terms of action score. This configuration proves to be the most efficient system overall for AP3's Non-IO tasks. Its dominance underscores the advanced design of the H100, optimized for parallel processing, enabling it to achieve lower runtimes and energy consumption compared to its counterparts. Additionally, the A100 configuration exhibits strong

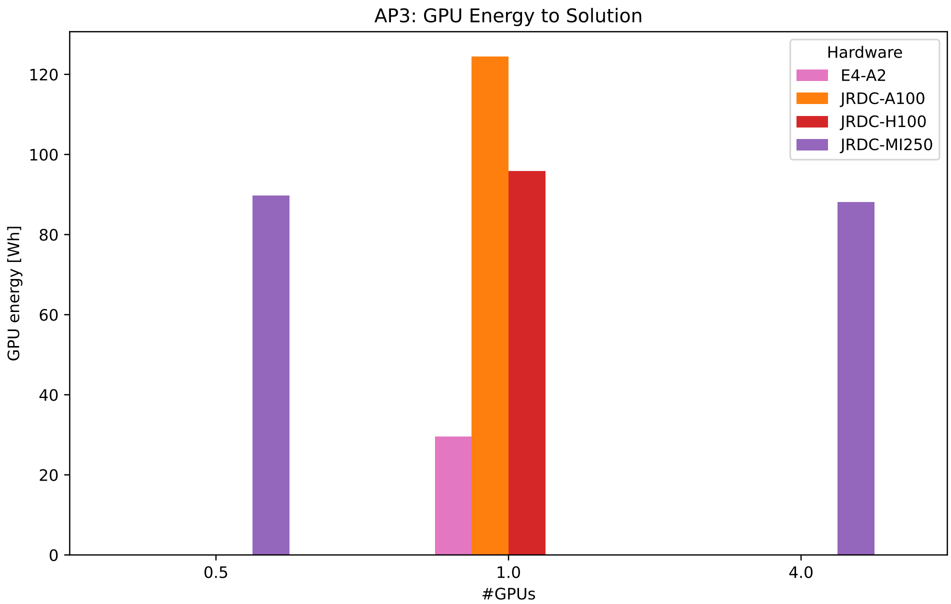


Figure 13: AP3 GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used during the Training phase. ap3-tr-best-gpu-en

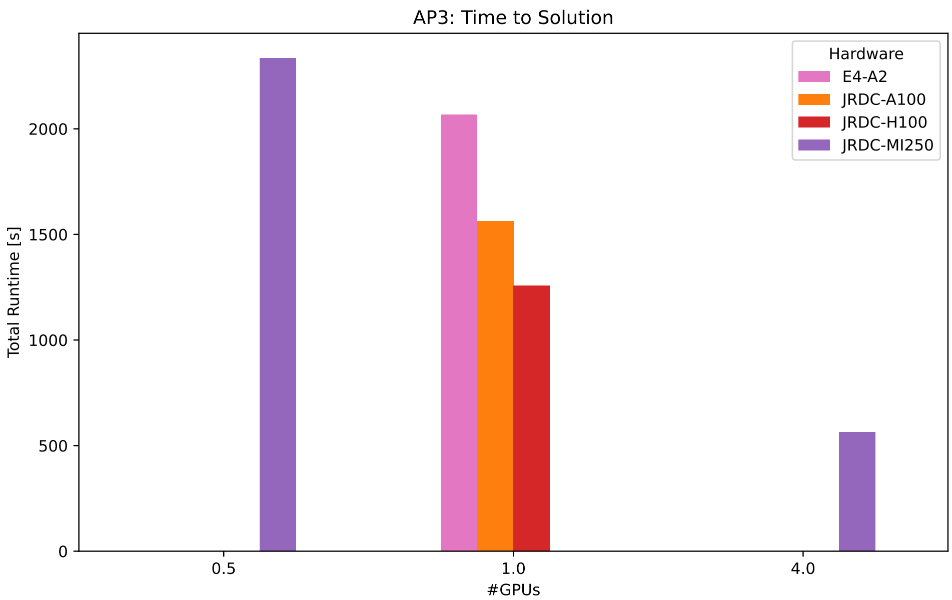


Figure 14: AP3 Time to solution: Time to solution considering the different hardware configurations during the Training phase. ap3-tr-best-runtime

performance in the 4-GPU case, boasting an action score slightly higher than that of the H100 configuration. Moreover, the MI250 configuration reports an action score

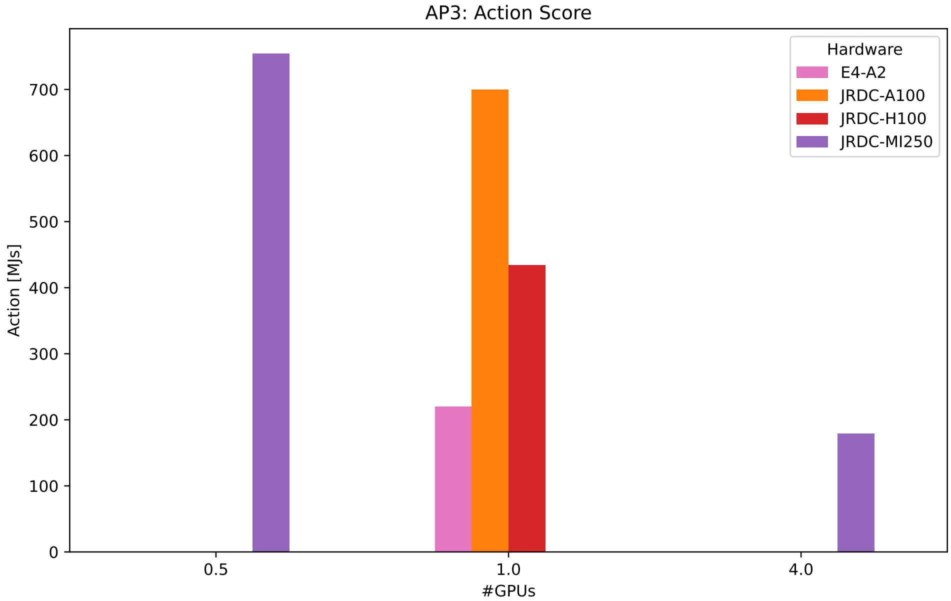


Figure 15: AP3 Action Score: Action Score considering the different hardware configurations used to determine the most performing one during the Training phase. ap3-tr-best-action

comparable to that of the single GPU Grace Hopper.
In contrast, the performance of the Graphcore IPU presents an intriguing deviation. Despite leading in Non-IO tasks for AP1, the IPUs demonstrate reduced efficiency and optimization for AP3, resulting in higher consumption and longer runtimes.
In summary, this concise overview highlights the JRDC-H100 with 4 GPUs as the optimal choice for Non-IO tasks in AP3, followed by the A100 and MI250 configurations. The E4-GH200 emerges as the most efficient option for single GPU setups.

3.3.3 Inference

The inference performance data for AP3 provides valuable insights into how different hardware configurations handle inference tasks. This analysis completes our exploration of AP3 across the full pipeline, Non-IO, and inference tasks, offering a comprehensive understanding of hardware performance across various computational scenarios.
From the plots reported in Fig. 19, 20 and 21, is notable how E4-A2 configuration demonstrates a good efficiency in inference tasks, showing the lowest action score of 0.078 MJs. This highlights the A2 GPU's ability to handle inference tasks with minimal energy consumption while maintaining competitive runtimes. The JRDC-A100 and JRDC-H100 configurations show close competition in inference performance,

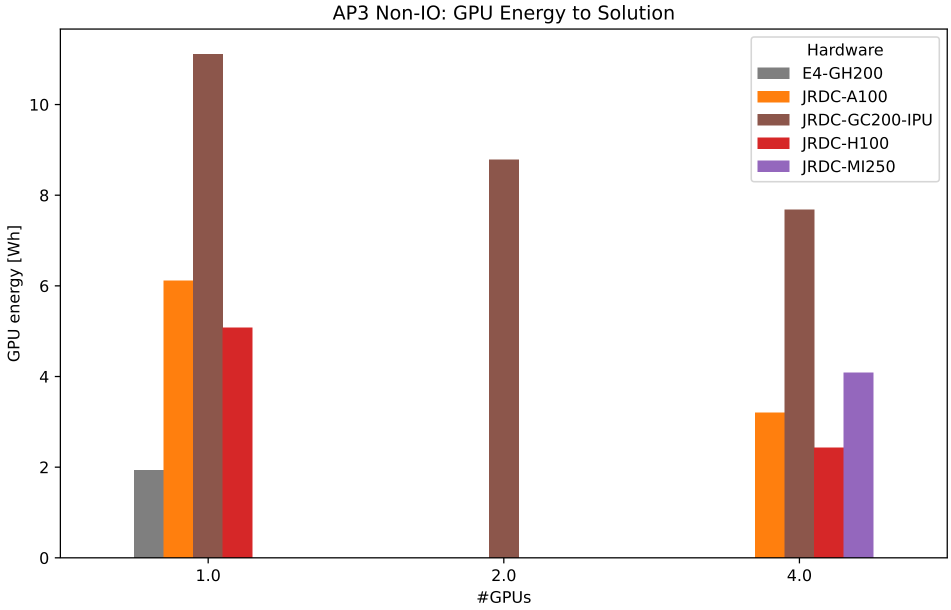


Figure 16: AP3 Non-IO GPU Energy to solution: GPU Energy to solution for Non-IO experiments considering the different hardware configurations used. ap3-nonio-best-gpu-en

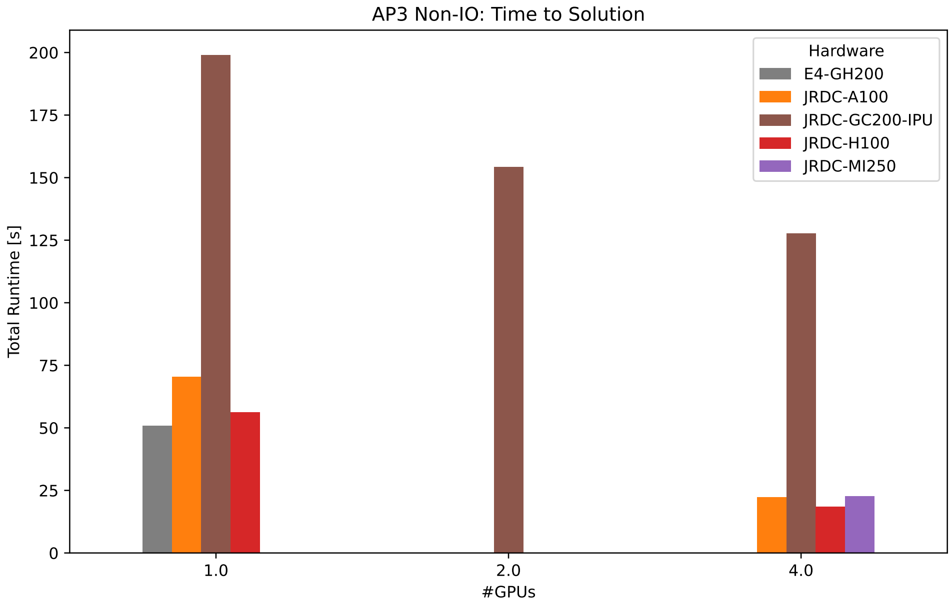


Figure 17: AP3 Non-IO Time to solution: Time to solution for Non-IO experiments considering the different hardware configurations used. ap3-nonio-best-runtime

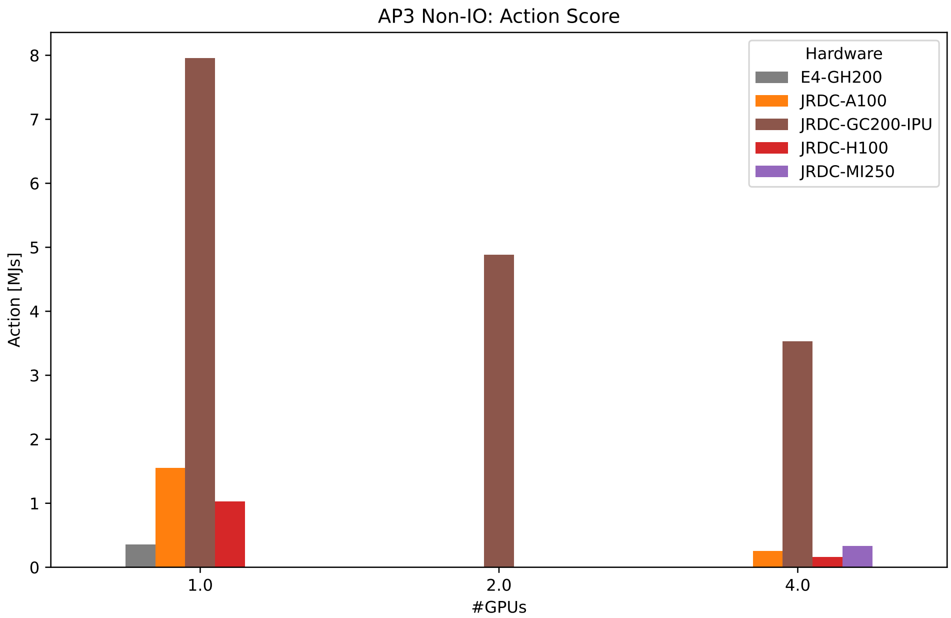


Figure 18: AP3 Non-IO Action Score: Action Score for Non-IO experiments considering the different hardware configurations used to determine the most performing one. ap3-nonio-best-action

with the H100 slightly leading in terms of a lower action score. However, both configurations exhibit higher energy consumption compared to the E4-A2, impacting their overall efficiency scores despite faster runtimes. Moreover, in a single GPU setup, the JRDC-MI250 delivers impressive efficiency, with an action score significantly lower than those of the A100 and H100 configurations. However, scaling up to 4 GPUs does not result in proportional efficiency gains, suggesting that multi-GPU configurations may not be optimal for inference tasks.

3.3.4 Host Power analysis

3.3.4.1 Full training

As seen in Fig. 22, for Application 3 we see an intersection for 1 GPUs where the H100 starts beating the A2 at a host power consumption of 300 Watt. Due to differing numbers of devices tested, comparing the performance of the MI250 is difficult.

3.3.4.2 Non-IO Experiments

As seen in Fig. 23, we see no intersections in the given power range. The best performer for 1 GPU is the GH200 and for 4 GPUs the H100.

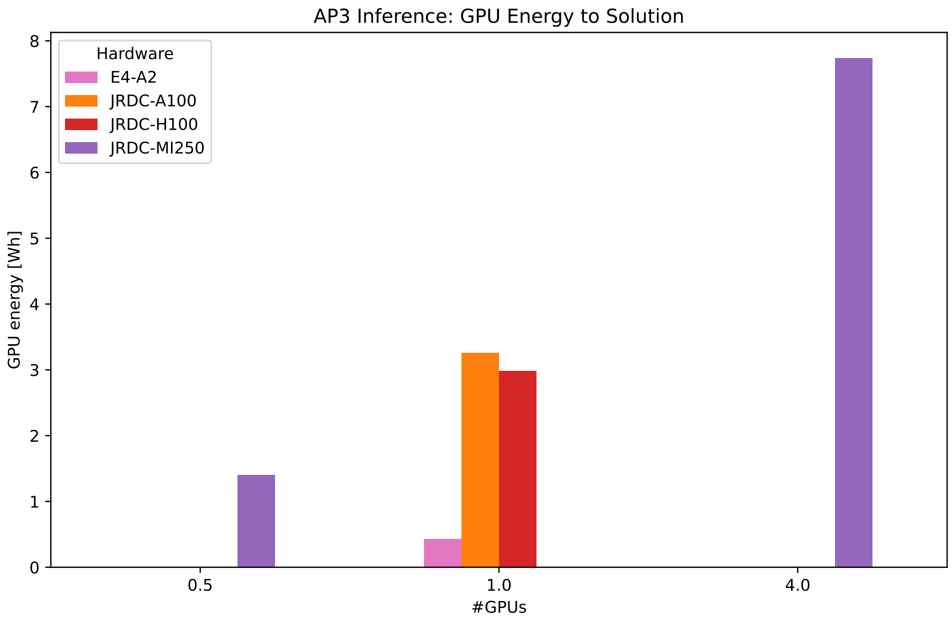


Figure 19: AP3 Inference GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used during the Inference phase. ap3-inf-best-gpu-en

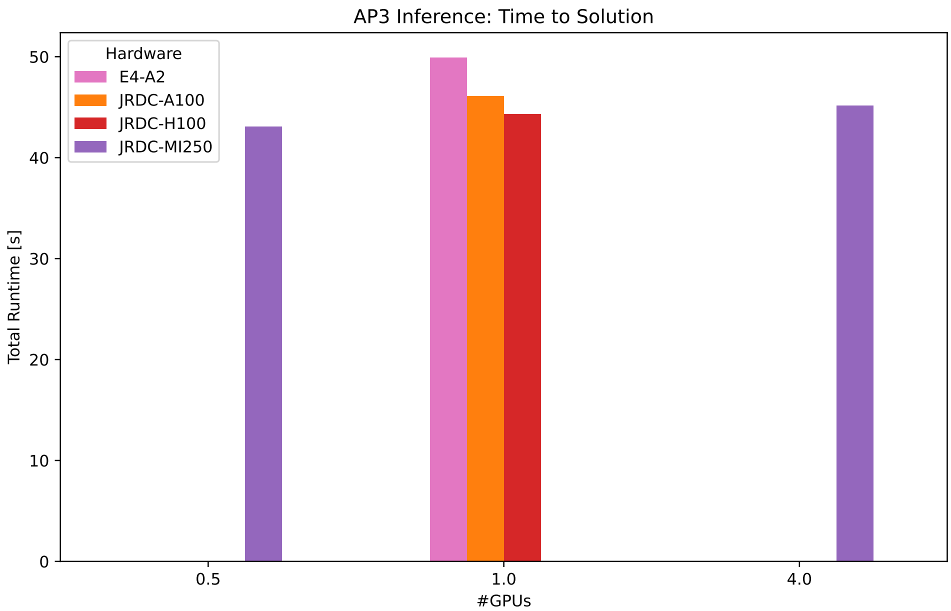


Figure 20: AP3 Inference Time to solution: Time to solution considering the different hardware configurations during the Inference phase. ap3-inf-best-runtime

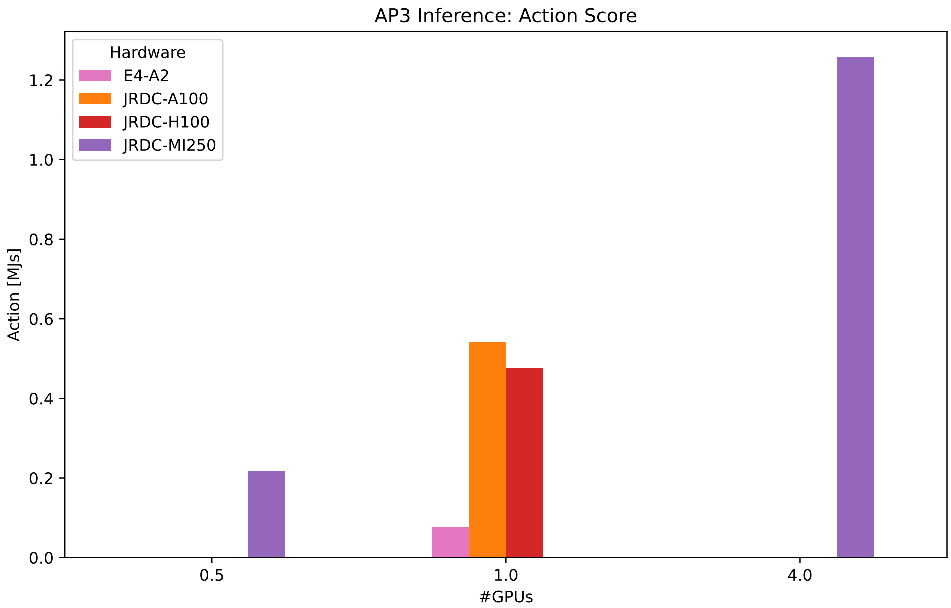


Figure 21: AP3 Inference Action Score: Action Score considering the different hardware configurations used to determine the most performing one during the Inference phase. ap3-inf-best-action

3.3.5 Conclusions

In this concluding section, we consolidate insights gathered from the training pipeline, Non-IO, and Inference experiments related to AP3.
Regarding the training pipeline, the top-performing systems are the 4-GPU JRDC-MI250 and the single GPU E4-A2 configurations. However, the dynamics shift in the Non-IO experiments, where multi-GPU systems featuring NVIDIA H100 and Grace Hopper come into play. Here, the E4-GH200 emerges as the most promising option in the single GPU configuration, while the JRDC-H100 configuration with 4 GPUs achieves the best overall performance. Furthermore, concerning Graphcore, based on the findings outlined here and in D3.7, we refrain from recommending the utilization of graphic IPUs except for AP1. Nevertheless, the system exhibits promise, and we recommend for further exploration into its potential.
In terms of inference, the E4-A2 stands out as the most efficient choice, striking a balance between runtime and energy consumption.

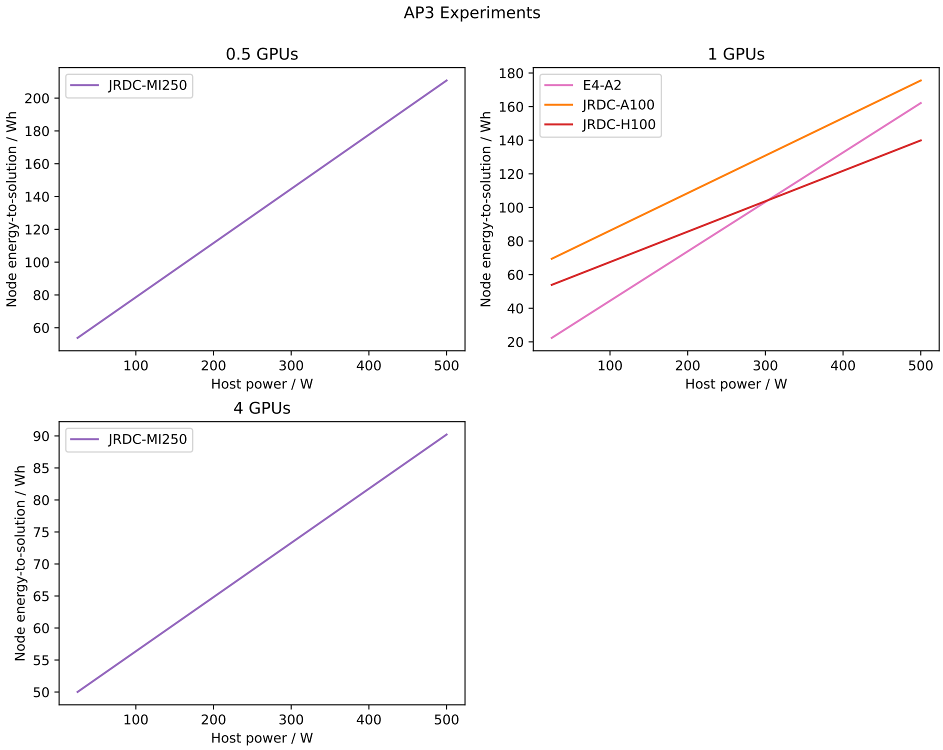


Figure 22: AP3 Energy-vs-Host-Power (Training): Full node energy consumption as a function of host power ap3-tr-etos

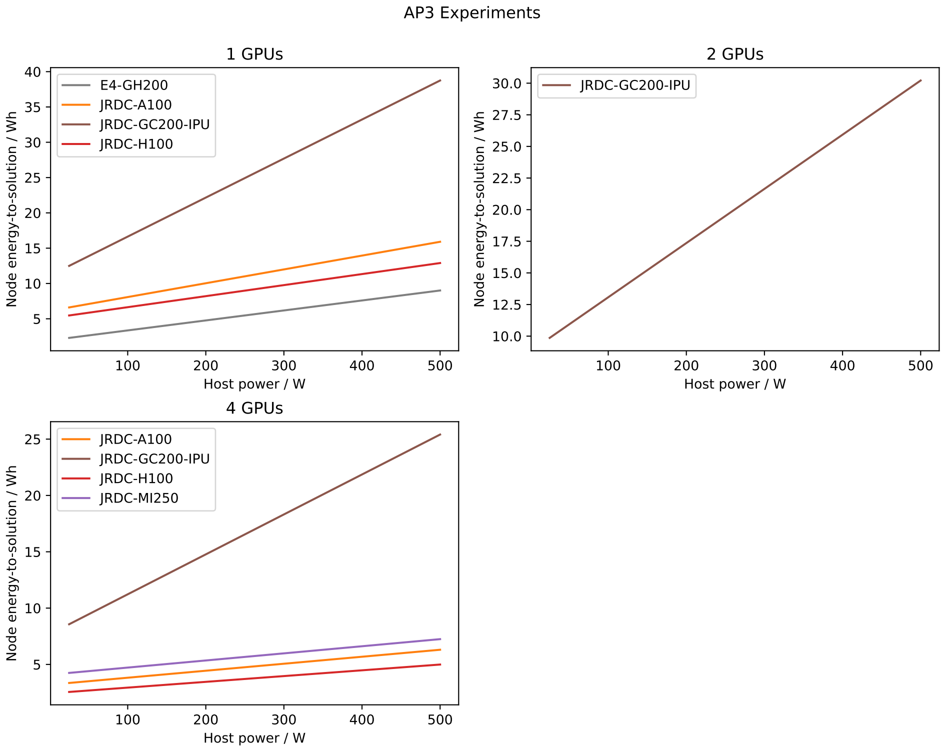


Figure 23: AP3 Energy-vs-Host-Power (Non-IO): Full node energy consumption as a function of host power `ap3-nonio-etos`

3.4 AP 4

For AP4, we're examining training performance without any parallelization, with each hardware configuration utilizing only one GPU, with the exception of the MI250 configuration, which uses only 1 GCD card (half of a GPU).

3.4.1 Full Training Pipeline

Upon examining the AP4 data, the E4-GH200 emerges as the most efficient among single GPU setups. It demonstrates an average power consumption (see Fig. 24) of around 122.34 Wh and boasts the shortest total runtime across all configurations (see Fig. 25). These observations are supported by its Action score of 1867.47 MJs, the lowest among all configurations (as reported in Fig. 26).
Meanwhile, the JRDC-A100 offer competitive alternatives, demonstrating its capacity to handle training workloads with respectable efficiency. Also the JRDC-MI250 gets quite good performance, even if it use only one of the two GDCs card at its disposal. While not leading the pack, the performance positions these two systems as viable options depending on specific task requirements and constraints.
However, the JRDC-H100 exhibits the highest consumption and a longer total runtime compared to the aforementioned configurations, suggesting it may not be as competitive in this context. Conversely, the E4-A2 showcases energy consumption comparable to the E4-GH200 and JRDC-A100 but with the longest runtime.

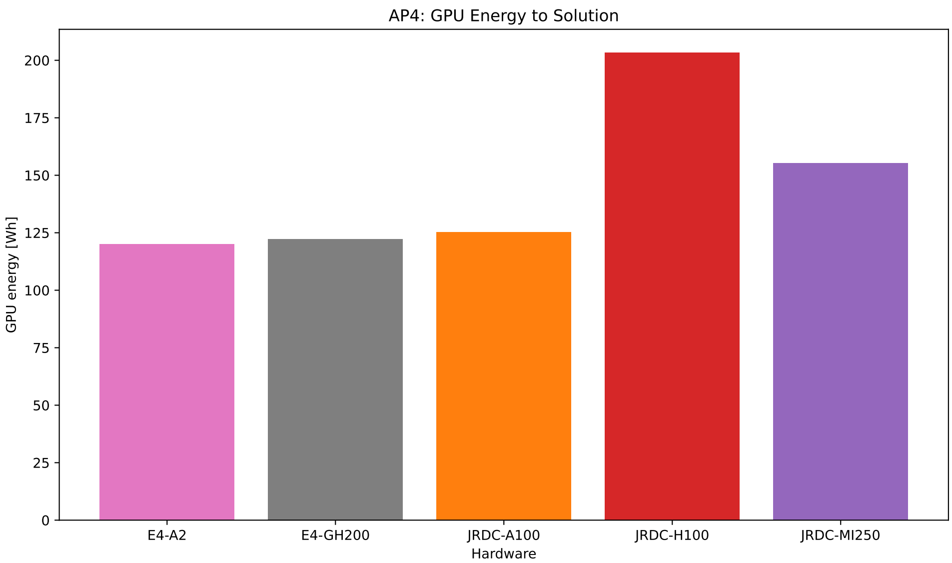


Figure 24: AP4 GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used during the Training phase. ap4-best-gpu-en

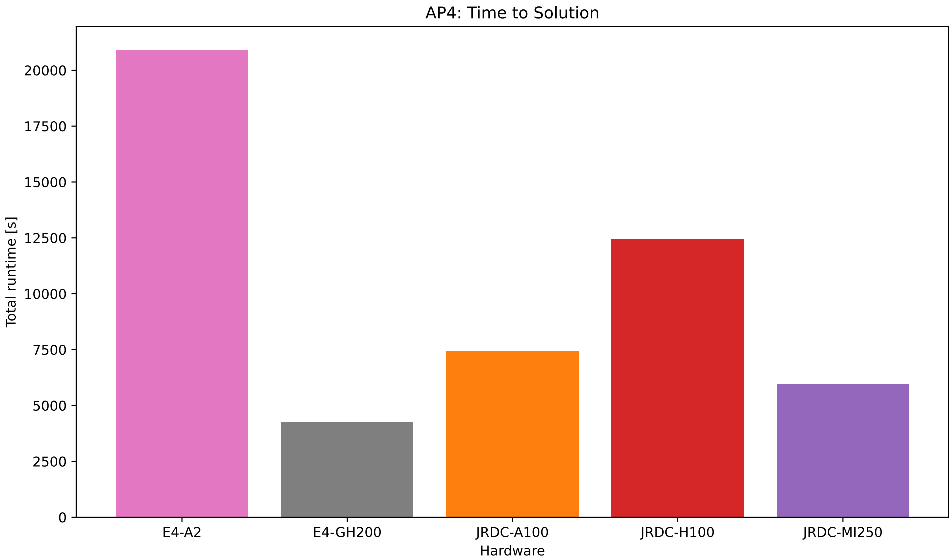


Figure 25: AP4 Time to solution: Time to solution considering the different hardware configurations during the Training phase. ap4-best-runtime

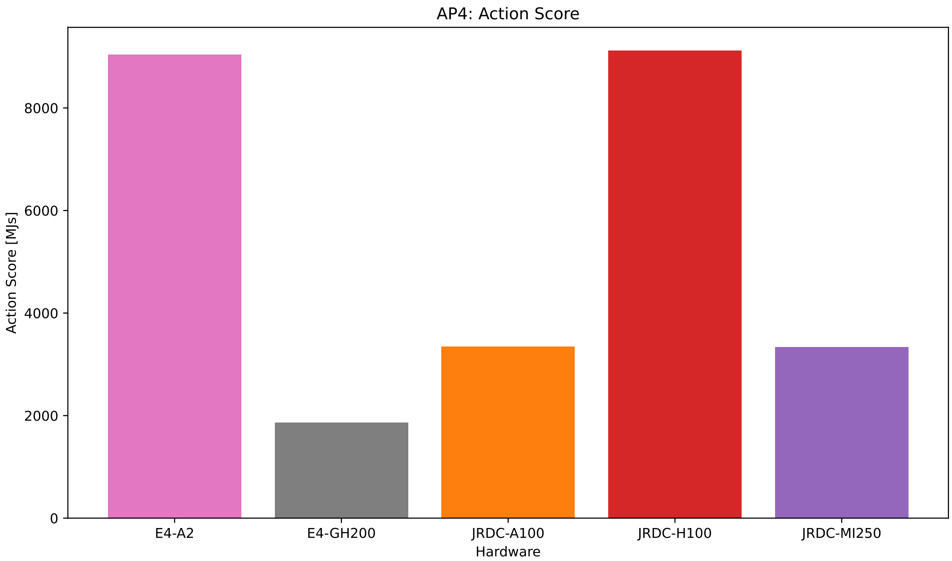


Figure 26: AP4 Action Score: Action Score considering the different hardware configurations used to determine the most performing one during the Training phase. ap4-best-action

3.4.2 Host Power analysis

As seen in Fig. 27, Application 4 did not use multi-GPU benchmarks. In case of the MI250 this means that only half of the device is used, making it difficult to compare

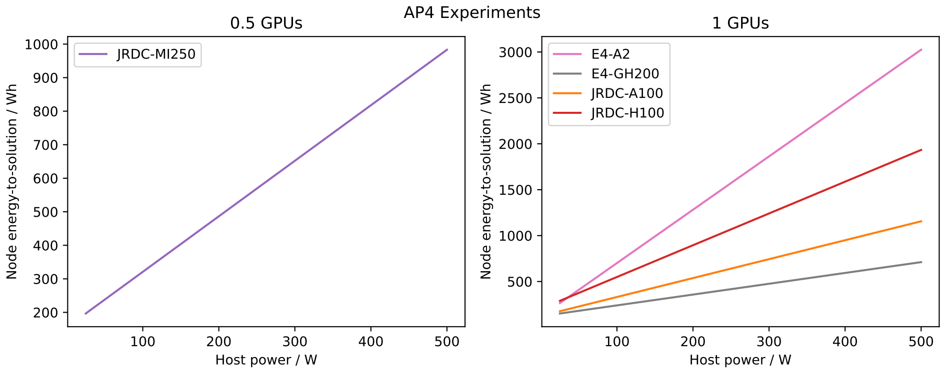


Figure 27: AP4 Energy-vs-Host-Power: Full node energy consumption as a function of host power ap4-tr-etos

to the other devices. The overall best performer is the GH200.

3.4.3 Conclusions

The analysis of AP4's training data underscores the E4-GH200 as the standout performer in efficiency among single GPU setups. With optimal balance in both speed and energy utilization, the GH200 demonstrates its ability in training scenarios where efficiency is critical.

3.5 AP 5

In the scope of AP5, we provide data concerning the full training pipeline and non-IO experiments. These will be elaborated upon in the following sections.

3.5.1 Full Training Pipeline

In evaluating the training pipeline data for AP5, as reported in Fig. 30, the E4-GH200 configuration is the standout for efficiency, achieving the lowest action score among the setups tested. On the other hand, JRDC configurations demonstrate significant improvements in runtime (refer to Fig. 29) and action as the number of GPUs increases, demonstrating commendable system scalability. This trend is particularly evident in the four-GPU configuration of JRDC-A100 and also the 4 GPUs JRDC-MI250, where both runtime and action score decrease significantly, also reporting a slight reduction in power consumption with the addition of involved GPUs (as reported in Fig. 28).

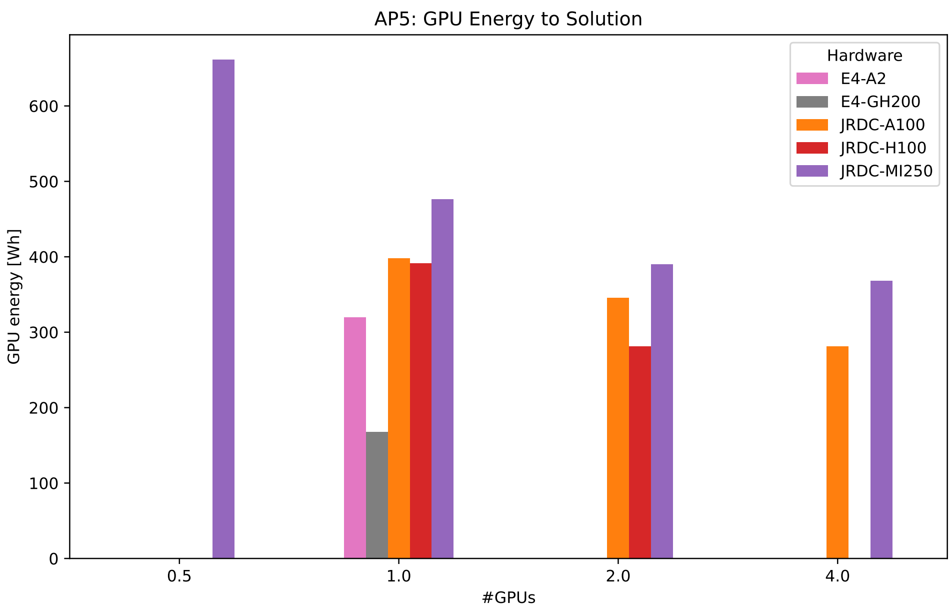
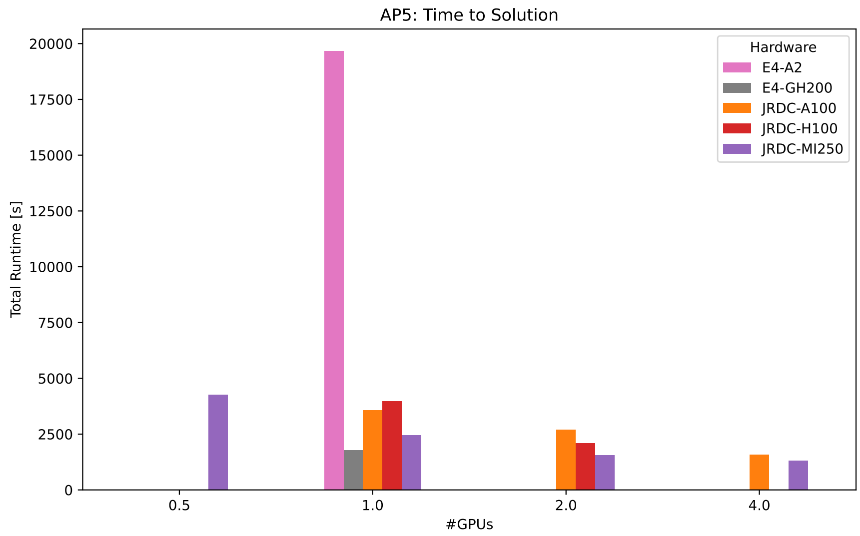


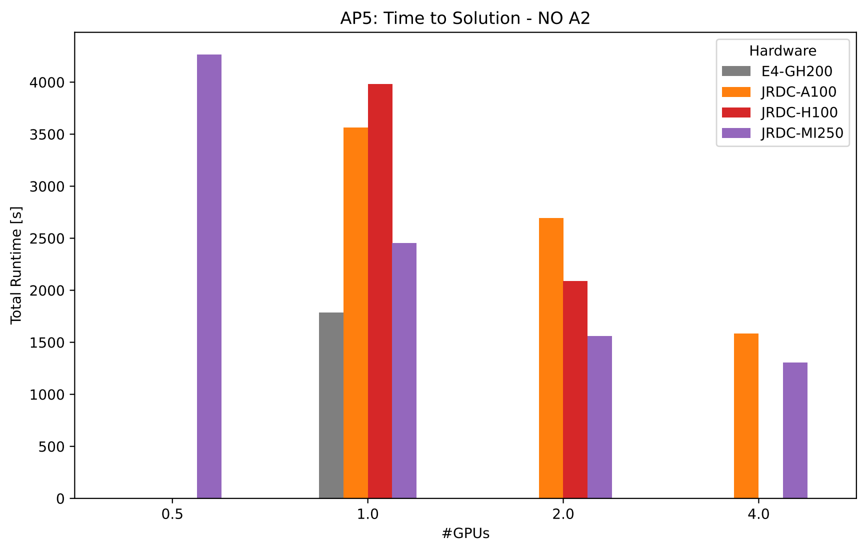
Figure 28: AP5 GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used. ap5-best-gpu-en

3.5.2 Non-IO Experiments

We delve now into the efficiency and performance of the various hardware configurations used by AP5 for the Non-IO experiments. Plots related to the metrics are reported in Fig. 31, 32 and 33.



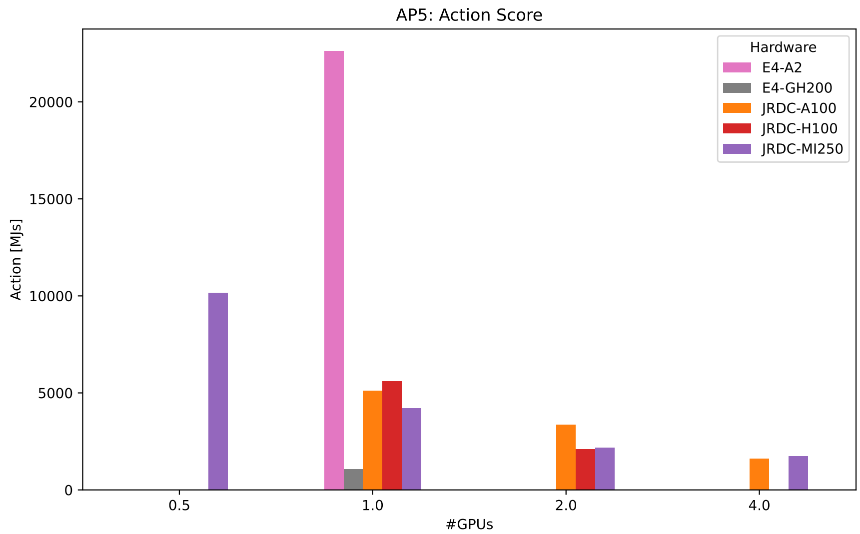
(a) Runtime comparison



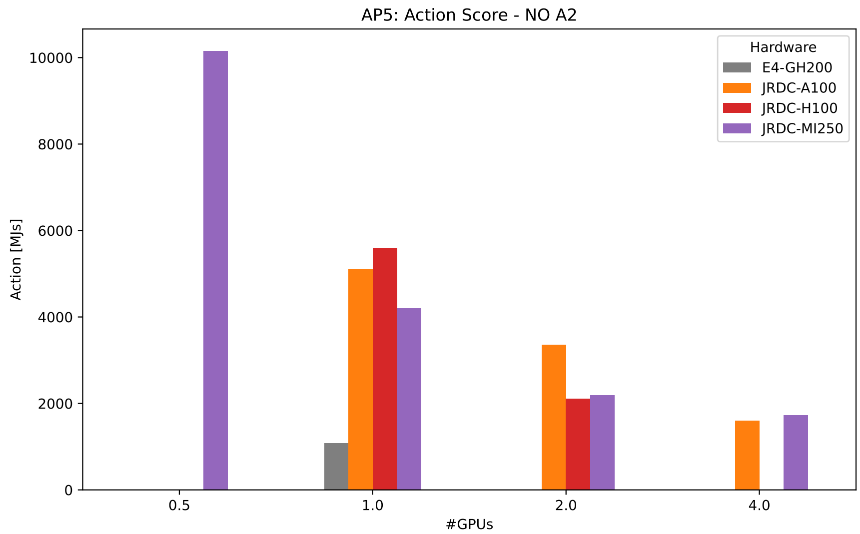
(b) Runtime comparison without A2

Figure 29: AP5 Time to solution: Time to solution considering the different hardware configurations used. ap5-best-runtime

The depicted data highlights the JRDC-MI250 with 4 GPUs as achieving the lowest action score among the configurations, marking it as the most efficient setup tested. With a score of 549.26, this configuration demonstrates a significant advantage in handling Non-IO tasks efficiently, showcasing effective use of multiple GPUs to optimize both runtime and energy consumption.



(a) Action comparison



(b) Action comparison without A2

Figure 30: AP5 Action Score: Action Score considering the different hardware configurations used to determine the most performing one. ap5-best-action

The JRDC-A100 configurations reveal an interesting trend; as the number of GPUs increases, the action score improves, indicating better efficiency. The transition from a single GPU to four GPUs sees a notable reduction in both in the total runtime and in the energy consumption, which reflect on the action scores. This suggests

that A100 GPUs benefit from parallel processing, enhancing their performance in Non-IO tasks.
E4-GH200 also achieves performance, with an action score of 1054.97 MJs. Despite being a single GPU setup, its efficiency is noteworthy, performing well in Non-IO tasks with relatively low energy consumption and faster runtime compared to other single GPU configurations.
E4-A2 and JRDC-MI250 in single GPU setups experience higher action scores, reflecting less efficiency in these particular tasks compared to their multi-GPU or more optimized counterparts. Particularly, the E4-A2's higher action score underscores its limited efficiency in this context, despite the potential for lower energy consumption with respect to the A100 and MI250 single GPU configurations.

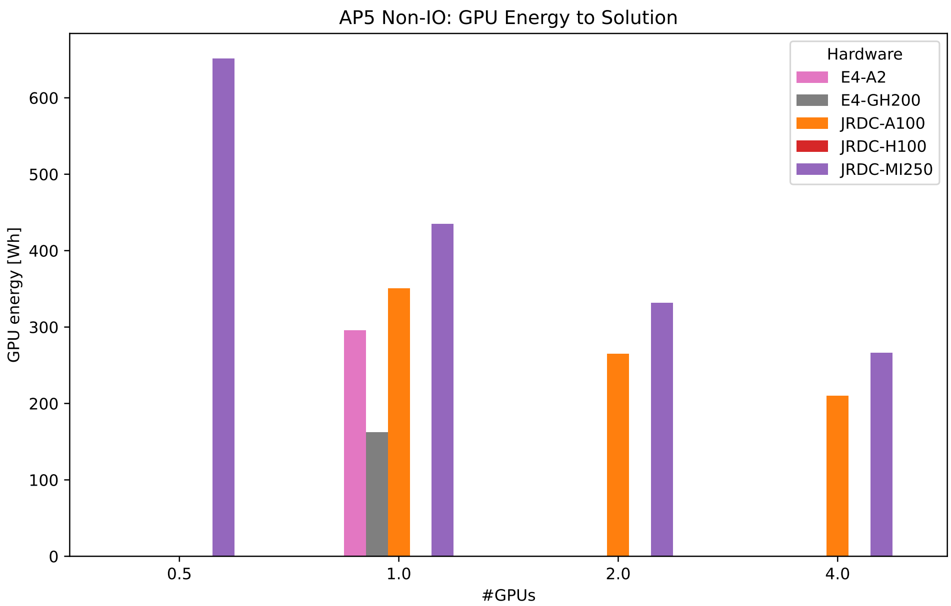
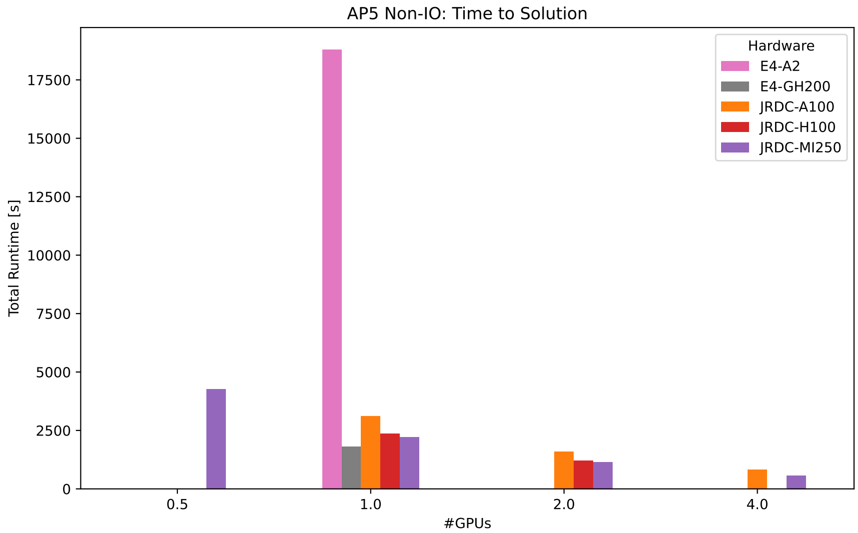


Figure 31: AP5 Non-IO GPU Energy to solution: GPU Energy to solution for Non-IO experiments considering the different hardware configurations used. ap5-nonio-best-gpu-en

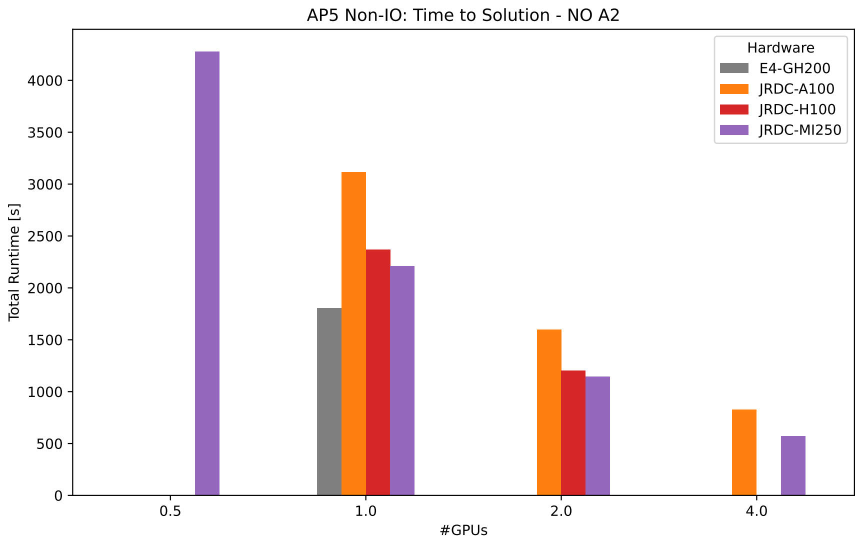
3.5.3 Host Power analysis

3.5.3.1 Full training

As seen in Fig. 34, we see some interesting results for Application 5. The overall best performer with 1 GPU is by far the GH200, for 2 GPUs it's the H100 and for 4 GPUs the A100. It is of note that this benchmark had suffered from temporary performance regressions on the H100 node. Another interesting point is that for 1



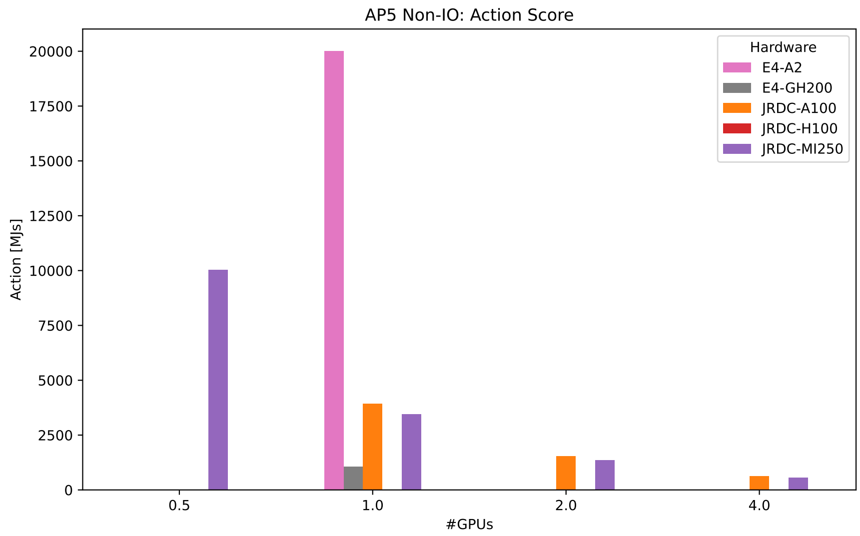
(a) Runtime comparison



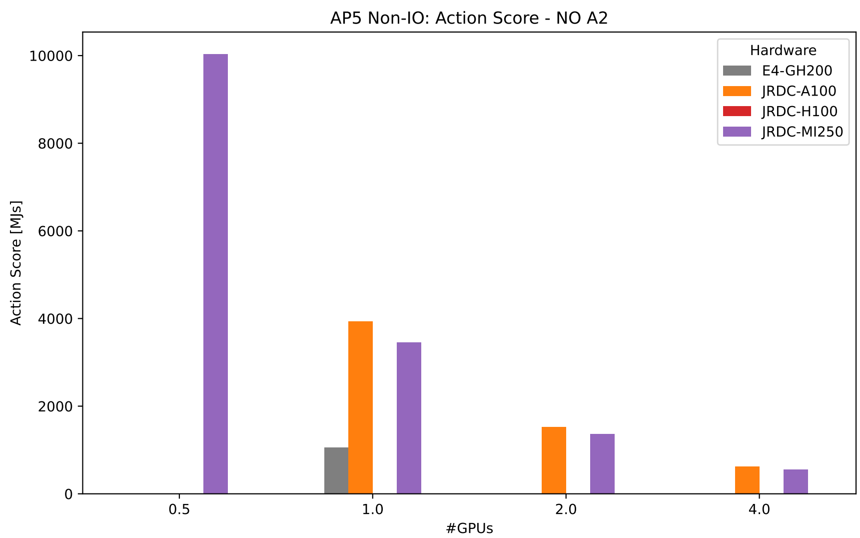
(b) Runtime comparison without A2

Figure 32: AP5 Non-IO Time to solution: Time to solution for Non-IO experiments considering the different hardware configurations used. ap5-best-nonio-runtime

GPU the MI250 starts outperforming the A100 and H100 at around 200 Watt host power and for 2 GPUs outperforms the A100 at 150 Watt host power.



(a) Action comparison



(b) Action comparison without A2

Figure 33: AP5 Non-IO Action Score: Action Score for Non-IO experiments considering the different hardware configurations used to determine the most performing one. ap5-nonio-best-action

3.5.3.2 Non-IO Experiments

As seen in Fig. 35, The best performer for 1 GPU is the GH200. We see the MI250 outperform the A100 above 350 Watt host power for 1 GPU and get very close to it at 500 Watt for 2 GPU. for 4 GPUs the A100 always outperforms the MI250.

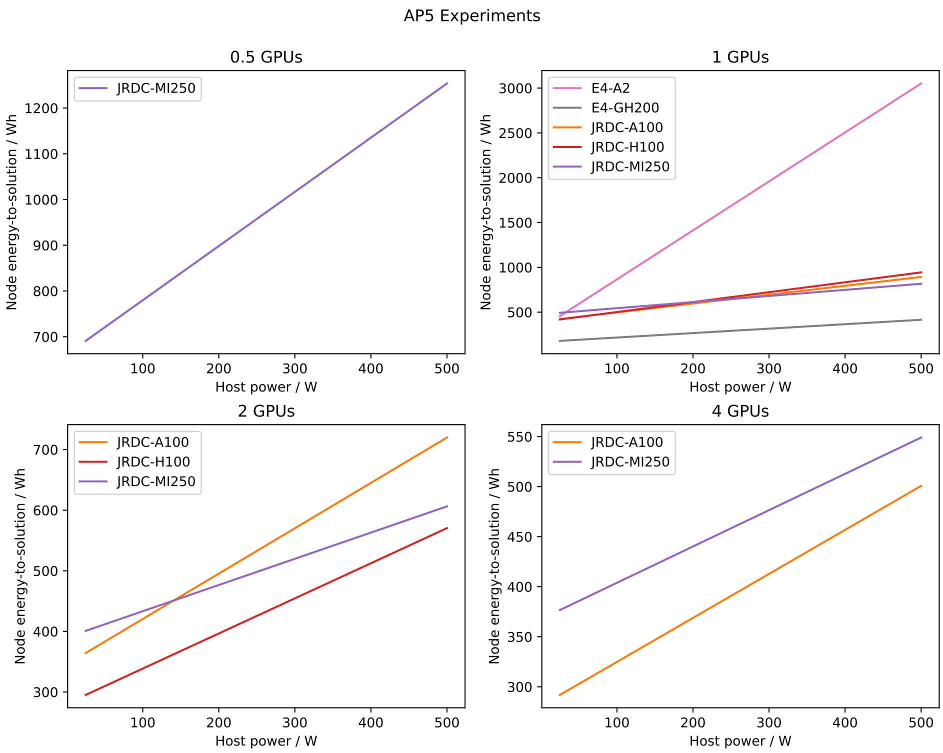


Figure 34: AP5 Energy-vs-Host-Power (Training): Full node energy consumption as a function of host power ap5-tr-etos

3.5.4 Conclusions

In this section, we summarize the findings from AP5 concerning both the training pipeline and Non-IO experiments.
In the training pipeline analysis, the top-performing system is the E4-GH200, followed by JRDC-A100 (4 GPUs) and MI250 (4 GPUs). However, the dynamics change slightly in the Non-IO experiments: the most performant system is represented by the 8-GCDs MI250 configuration, followed by the 4-GPU JRDC-A100 and the E4-GH200.

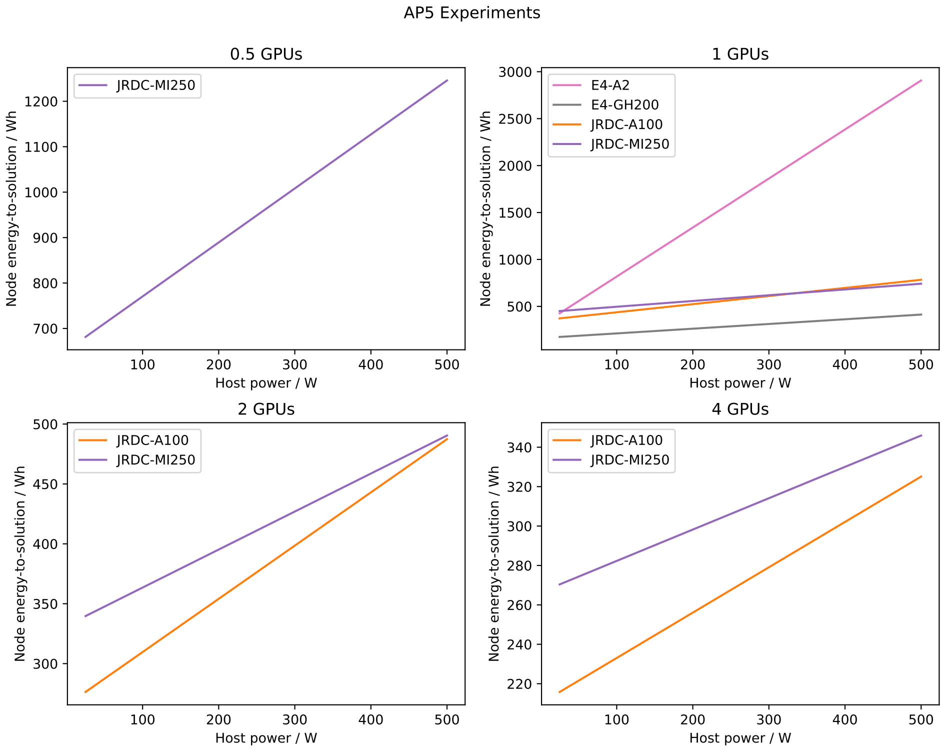


Figure 35: AP5 Energy-vs-Host-Power (Non-IO): Full node energy consumption as a function of host power ap5-nonio-etos

3.6 AP 6

Our analysis of applications concludes with AP6, which focuses on the training pipeline. This application explores configurations that not only vary in the number of GPUs but also in the number of nodes involved.

3.6.1 Full Training Pipeline

The E4-A2 configurations show a clear trend of increased efficiency with more GPUs. Starting from a single GPU to four GPUs (two GPUs for each node), there's a noticeable decrease in both total runtime and action score (see Fig. 36, 37 and 38), indicating that E4-A2 scales well for this particular training task. The action score improves significantly as the number of GPUs increases, highlighting effective parallel processing capabilities. The E4-GH200 configurations also show improvements with the introduction of multi-node parallelization, achieving a score second only to that of the E4-A2 configuration. Also, the JRDC-A100 configuration shows a moderate performance improvement with two and three GPUs, as evidenced by the action scores, passing from 5491.58 to 2509.34 MJs. Then, JRDC-MI250 and JRDC-H100 exhibits increased action scores with more GPUs, showing that adding GPUs leads to higher energy consumption without proportional gains in runtime efficiency, especially notable in the JRDC-H100's two and four GPU setup showing a significant jump in action score.

3.6.2 Conclusions

3.6.3 Host Power analysis

Several caveats apply to the results seen in Fig. 39 for Application 6. On JSC machines (JRDC-A100, JRDC-H100 and JRDC-MI250) an issue with the launch script configuration prevented the application from utilizing multiple devices, leading to negative scaling with the number of devices. The configuration was fixed as of this deliverable, but reruns of the benchmarks could only be performed for the A100 device, the results for the H100 and MI250 are still affected. Furthermore, the per-node parallelism was limited to 3 devices due to the high memory consumption of the application.
We see that for 1 GPU the best 3 performers reverse order at a host power of 225 Watt. For lower host powers the A2 comes in first place, followed by the GH200 and A100. For higher power the A100 performs best, followed by the GH200 and A2.
For 2 GPUs the best performers are the GH200 and A2, with the GH200 overtaking the A2 at a host power of 375 Watt.

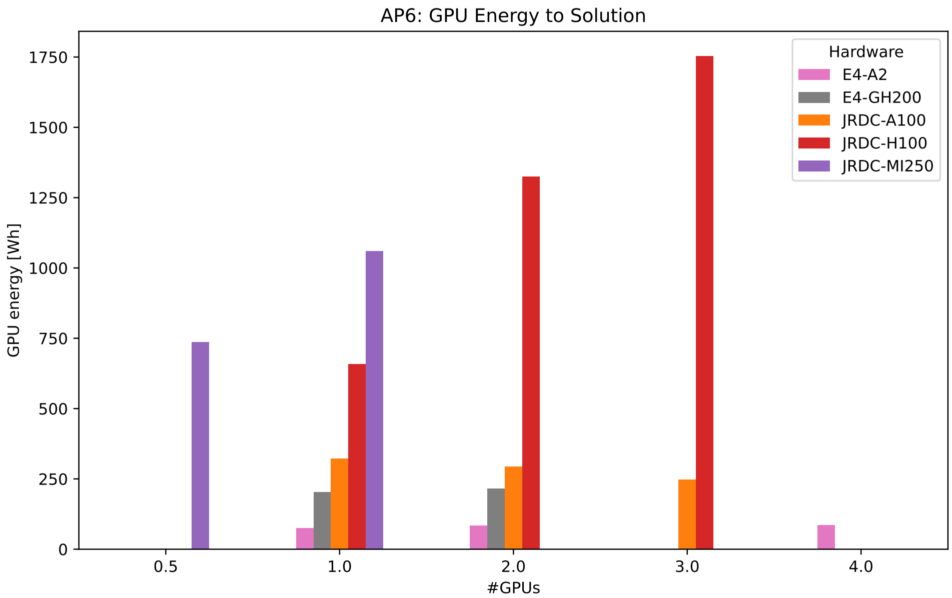


Figure 36: AP6 GPU Energy to solution: GPU Energy to solution considering the different hardware configurations used during the Training phase. ap6-best-gpu-en

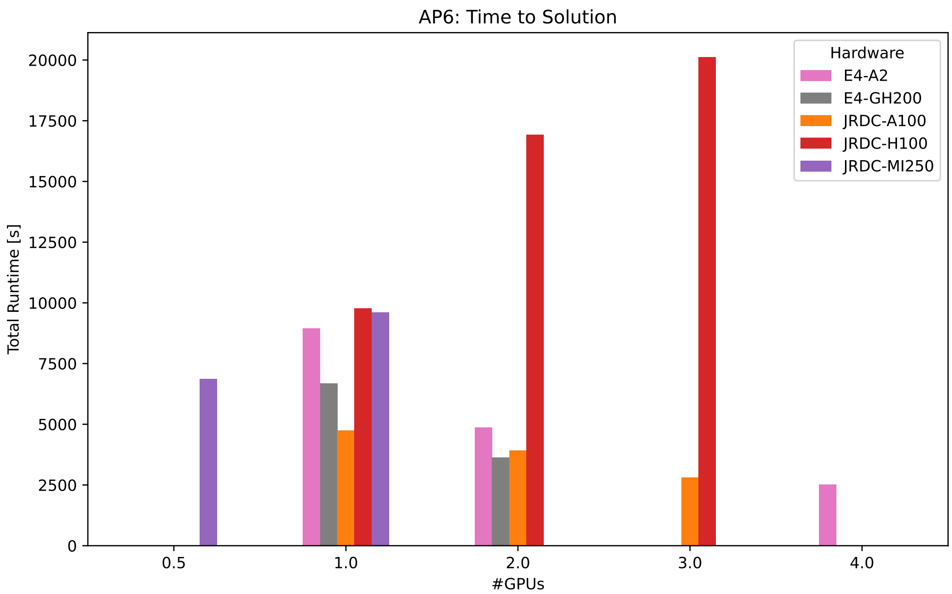


Figure 37: AP6 Time to solution: Time to solution considering the different hardware configurations during the Training phase. ap6-best-runtime

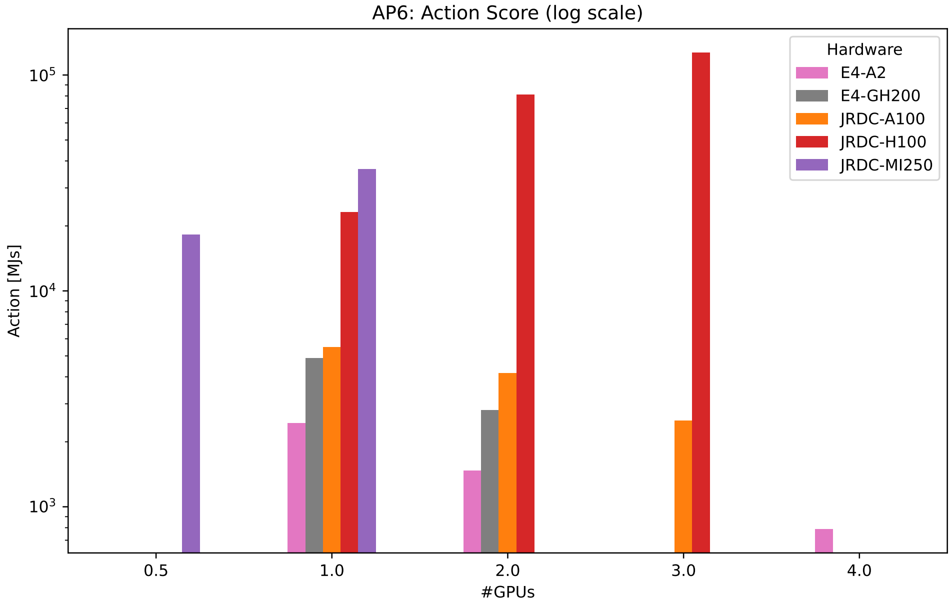


Figure 38: AP6 Action Score: Action Score considering the different hardware configurations used to determine the most performing one during the Training phase. ap6-best-action

3.6.4 Conclusions

The analysis of AP6's training pipeline reveals that the E4-A2 configurations are particularly efficient, with noticeable improvements in performance and energy use as more GPUs and nodes are added. Also E4-GH200 and JRDC-A100 also performed well, while JRDC-H100, and JRDC-MI250 configurations demonstrate varying degrees of efficiency, with generally higher action scores that indicate less optimal energy consumption relative to performance gains.

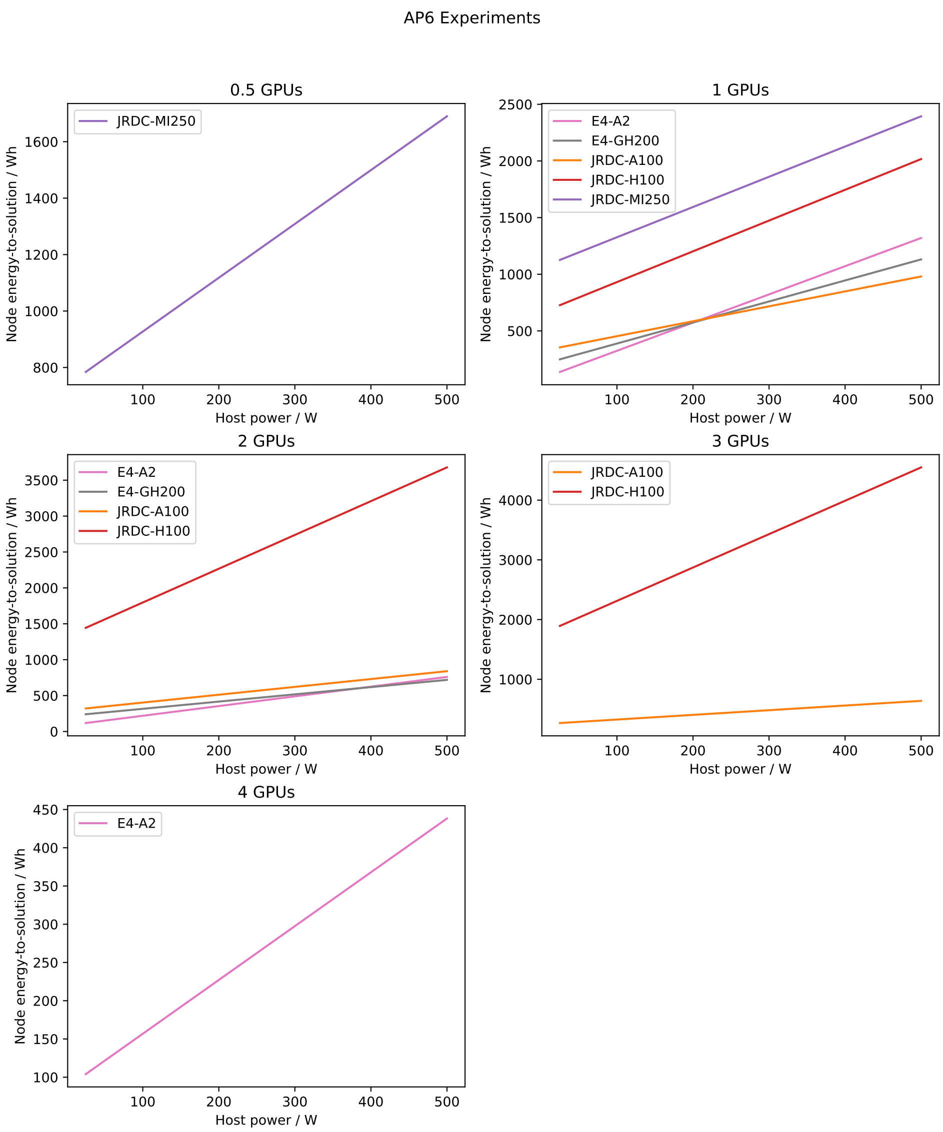


Figure 39: AP6 Energy-vs-Host-Power: Full node energy consumption as a function of host power ap6-tr-etos

3.7 Comparative Analysis

In this comparative analysis of hardware configurations, we examine the performance of various hardware setups across all six distinct applications outlined within the MAELSTROM project. Our objective is to identify configurations that demonstrate optimal performance across the board. To facilitate a comprehensive comparison, we've compiled a table summarizing the top-performing hardware configurations identified across these applications. Each application presents unique computational challenges, emphasizing the importance of this comparison in highlighting the versatility and adaptability of different hardware solutions. Table 2 displays the top three best-performing hardware setups overall, aiding in the selection of configurations that best align with the requirements and constraints across all applications.

Table 2: Comparative Analysis of Hardware Configurations Across Applications. The number of GPUs in the configuration is shown between brackets

Application	Top Performer	Second Best	Third Best
AP1	E4-GH200 (1)	JRDC-H100 (4)	-
AP1-NonIO	JRDC-H100 (4)	JRDC-GC200-IPU (4)	JRDC-A100 (4) / E4-GH200 (1)
AP2	JRDC-H100 (1)	JRDC-A100 (1)	E4-GH200 (1)
AP3	JRDC-MI250 (4)	E4-A2 (1)	JRDC-H100 (1)
AP3-NonIO	JRDC-H100 (4)	JRDC-A100 (4)	JRDC-MI250 (4) / E4-GH200 (1)
AP3-Inf	E4-A2 (1 GPU)	JRDC-MI250 (1)	-
AP4	E4-GH200 (1)	JRDC-MI250 (1)	JRDC-A100 (1)
AP5	E4-GH200 (1)	JRDC-A100 (4)	JRDC-MI250 (4)
AP5-NonIO	JRDC-MI250 (4)	JRDC-A100 (4)	E4-GH200 (1)
AP6	E4-A2 (All)	JRDC-A100 (3)	E4-GH200 (2)

The table identifies hardware configurations that consistently deliver top-tier performance across the applications. Our analysis, grounded in comparative performance data, has led us to identify two hardware configurations that stand out for their respective strengths in training and inference tasks: the GH200 and the NVIDIA A2. The GH200 configuration has demonstrated superior performance in training phases across various applications, competing effectively with multi-GPU configurations despite being a single-device setup. Its robust computational capabilities make it ideally suited for handling the complex, data-intensive models typical in weather and climate simulations. The GH200's efficiency and scalability ensure that as models grow in complexity and size, the infrastructure can adapt, providing faster iterations and enabling more comprehensive simulations. For the inference phase, the NVIDIA A2 GPU emerges as the most effective option. It offers the necessary speed and energy efficiency for running trained models, making it particularly well-suited for deploying W&C models in operational settings where rapid data processing is paramount.

4 Best-Performing Hardware Configuration

Achieving excellence in W&C applications requires an infrastructure that not only meets current computational demands but is also scalable and energy-efficient to support future advancements. Our analysis, based on comparative performance data, has led us to highlight the GH200 hardware configuration for its strengths in training across all the six proposed applications. We delve into the specifics of this high-performance architecture to showcase its ability to effectively manage W&C AI workloads, both in training and inference.

4.1 NVIDIA GH200

The NVIDIA GH200 Grace Hopper architecture [1] [2], reported in Fig. 40 and 41, combines the power of the NVIDIA Hopper GPU with the versatility of the NVIDIA Grace CPU. This fusion is made possible by a high-bandwidth and memory-coherent NVIDIA NVLink Chip-2-Chip (C2C) interconnect within a single Superchip, alongside support for the new NVIDIA NVLink Switch System.

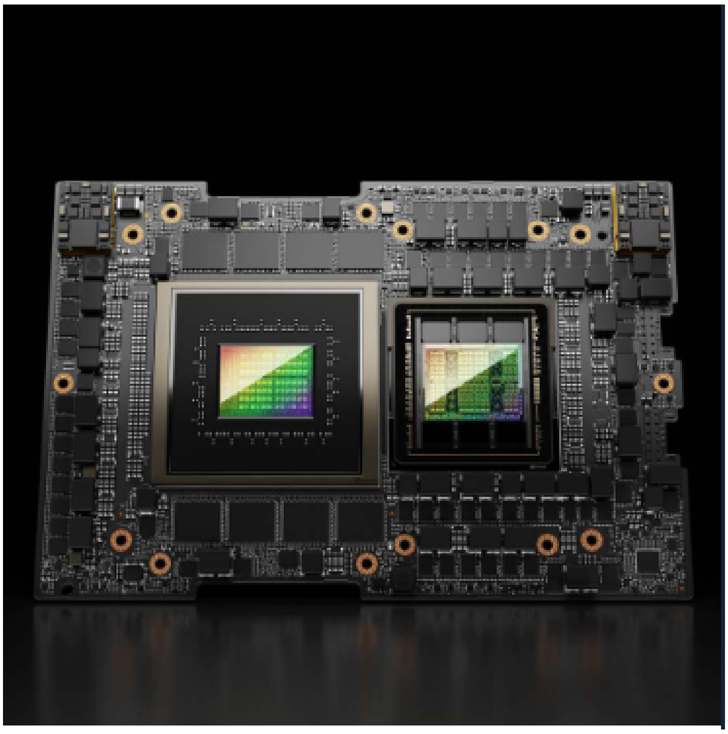


Figure 40: NVIDIA GH200: Grace Hopper Superchip *gh200*

The Grace CPU integrates 72 Neoverse V2 Armv9 cores with up to 480GB of server-class LPDDR5X memory with ECC. In contrast to an eight-channel DDR5 design, the

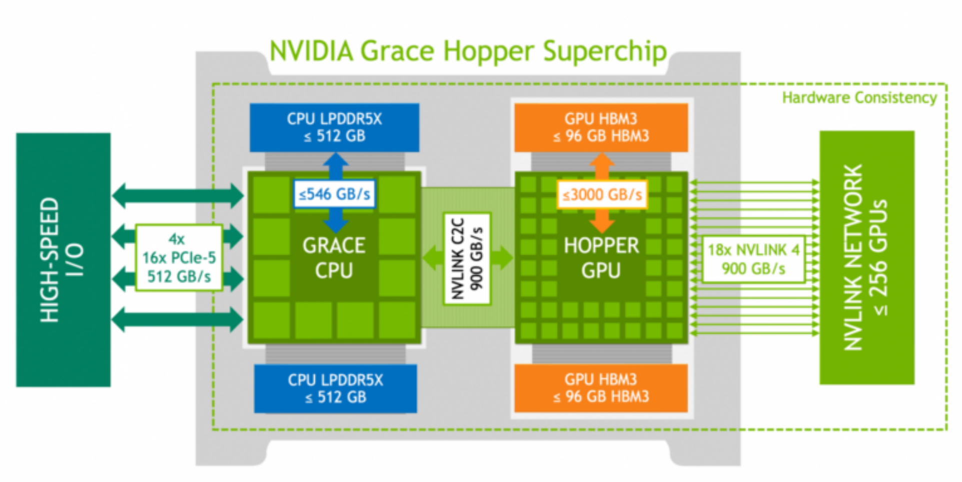


Figure 41: NVIDIA GH200: Grace Hopper logical overview. gh200-log

Grace CPU LPDDR5X memory subsystem offers up to 53 percent more bandwidth at one-eighth the power per gigabyte per second.
The H100 Tensor Core GPU is the latest data center GPU by NVIDIA, Powered by the innovative Hopper GPU architecture, the H100 introduces several key advancements. Firstly, it features fourth-generation Tensor Cores, which excel at accelerating matrix computations across a wide range of AI and HPC tasks. Additionally, the integration of a new Transformer Engine enables the H100 to achieve speedups in AI training, up to 9 times faster than the previous generation, and up to 30 times faster AI inference. Furthermore, the H100 introduces Secure Multi-Instance GPU (MIG) technology, allowing the GPU to be partitioned into isolated instances tailored to different workloads.
GH200 Superchip incorporates HBM3 memory, utilizing 96GB and delivering 4TB/s of memory bandwidth. The HBM is seamlessly integrated with CPU memory via NVLink-C2C, providing up to 624GB of fast-access memory to the GPU.
NVLink-C2C serves as NVIDIAs memory-coherent, high-bandwidth, and low-latency interconnect for superchips, serving as the backbone of the GH200. It delivers up to 900GB/s total bandwidth. This memory coherency allows concurrent and transparent access to both CPU and GPU resident memory, freeing developers from explicit memory management and enabling them to focus on algorithms. NVLink-C2C also enables applications to oversubscribe GPU memory and utilize NVIDIA Grace CPU memory at high bandwidth. With up to 480GB of LPDDR5X CPU memory per Grace Hopper Superchip, the GPU gains direct high-bandwidth access to an additional 480GB of memory.

5 Conclusion

The culmination of our efforts in previous benchmarking analyses is encapsulated within this deliverable, aimed at delineating a solution design tailored for W&C applications. Throughout the project's duration, we have observed substantial advancements in both hardware technology and the applications themselves, indicative of a visible evolution in our approach.

Within this deliverable, we revisited and analyzed the hardware configurations employed in the last benchmarking phase. Our objective was to scrutinize their capabilities and optimizations in executing machine learning benchmarks pertinent to the W&C domain.

Our process commenced with an examination of the top-performing hardware configurations for each application, with the goal of pinpointing the most effective and efficient solutions to meet our computational requirements. Subsequently, a comparative analysis was undertaken to identify hardware configurations that not only excel in individual applications but also strike a balance across all evaluated software specifically tailored for W&C needs.

During our examinations of distributed File Systems (FS) over the network (e.g. Ethernet, Infiniband), detailed in D3.6, we noticed a significant performance boost when repeatedly accessing the same dataset, particularly when the server cache was activated. This enhancement results from the dataset being fully cached in the server's memory after the initial read.

Moreover, the utilization of high-performance local storage (NVMe) for dataset reads proved to be even more impactful in enhancing performance compared to distributed FS.

Following our extensive analysis, we have selected the GH200 hardware configuration for its outstanding training performance across all six proposed applications. This conclusion is elaborated upon in the final chapter of this document. While the defined architecture offers limited customization options, the inclusion of an NVMe space tailored to accommodate datasets could be a valuable consideration.

The NVIDIA GH200 Grace Hopper superchip represents a pioneering advancement, marking the beginning of true heterogeneous accelerated platforms for high performance computing and AI workloads. Its integration of GPU and CPU strengths, coupled with a simplified and efficient heterogeneous programming model,positions it as as an innovative solution in the field.

References

[1] NVIDIA. Nvidia gh200 grace hopper superchip architecure: Performance and productivity for strong-scaling hpc and giant ai workloads. *https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper*.

[2] NVIDIA. Nvidia gh200 grace hopper superchip: The breakthrough processor for large-scale ai and high-performance computing (hpc) applications.

Document History

Version	Author(s)	Date	Changes
	Name (Organisation)	dd/mm/yyyy	
1.0	E4	28/03/2024	Final version

Internal Review History

Internal Reviewers	Date	Comments
Peter Dueben (ECMWF)	27/03/2024	Accepted with minor revisions
Saleh Ashkboos (ETH)	27/03/2024	Accepted with minor revisions