ISBN: 978-3-95450-247-9 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2024-MOPG18

INTEGRATING SUSTAINABLE COMPUTATIONAL STRATEGIES IN LIGHT SOURCE ACCELERATOR UPGRADES

P. Niknejadi*, S. Ackermann, E. Ferrari, T. Lang,
G. Paraskaki, L. Schaper, S. Schreiber, M. Vogt, J. Zemella
Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany
M. Asatrian, W. Hillert, F. Pannek, D. Samoilenko
Universität Hamburg Institute of Experimental Physics, Hamburg, Germany

Abstract

The operation of light source accelerators is a complex process that involves a combination of empirical and theoretical physics, simulations, and data-intensive methodologies. For example, the FLASH1 beamline at DESY is upgrading to an external seeding FEL light source. We utilize special diagnostics, machine learning algorithms, and comprehensive simulations to achieve this. To optimize resources, we constantly look to improve our approach, allowing us to robustly control the accelerator and meet the desired stability of our users. Machine learning and GPU-based algorithms have become crucial, enabling us to employ advanced optimization techniques and possibly Artificial Intelligence. However, in many cases, it is imperative to establish a robust mechanism for simulations involving large particle numbers to ensure that future upgrades and experiments can effectively and sustainably leverage these computational strategies.

EMERGING AI AND ML IN FELS

The recent "First Large Language Models in Physics Symposium" [1] coupled with the growing interest in harnessing Generative Machine Learning (ML) models for accelerator operation, data processing, and analysis in state-of-the-art light sources, hints at a future marked by a substantial integration of Artificial Intelligence (AI) in the fields of beam and accelerator physics as well as photon science. This integration is poised to surpass the currently advanced machine-learning applications and potentially extend these machines' capabilities. This evolving landscape sets the stage for the transformative impact of AI and computational strategies on the efficiency, sustainability, and performance of Free Electron Lasers (FELs). After highlighting six critical strategies for sustainability in the computational domain of accelerator and light source operation, this paper will discuss our experiences with implementing several practices and extract further insights relevant to the scientific community. We use the efforts made during the start-to-end simulation and optimization for the FLASH1 upgrade planning as a case study, highlighting the increasing resource demand due to advances in computing technology. This detailed examination reflects current progress and trends and opens a dialogue on future cultural adjustments in our field.

KEY STRATEGIES

As we continue to push the envelope in light source technologies and strive for improved performance, sustainable computational strategies become increasingly crucial in these fields. These strategies closely align with trends observed in AI and ML-driven industries, reflecting a broader shift towards more efficient, innovative, and adaptive technological practices. Our review of industry standards [2] and environmentally sustainable computational science [3] has identified key strategies that not only guide daily operations but also significantly influence broader initiatives such as facility design and upgrades. These strategies include:

- **Scalability and Modular Design**: Tailoring simulation and analysis tools to be adaptable across various challenges ensures that systems can grow and evolve without extensive overhauls.
- **Data-Driven Methodologies**: Implementing advanced analytics to enhance optimizations and performance, ensuring decisions are informed by robust data analysis.
- **High-Performance Computing (HPC) Utilization**: Employing HPC resources to not only bolster computational capacity but also optimize operational costs through energy-efficient scheduling.
- **Custom Software Solutions**: Developing custom software for parallel processing and memory management that is finely tuned to specific operational needs.
- **Regular Benchmarking, Validation, and Evaluation**: Establishing continuous evaluation protocols is pivotal in maintaining high reliability and refining our processes over time. Integral to these protocols is the quantitative assessment of our computational work's environmental impact, specifically focusing on evaluating the carbon footprint [4, 5].
- **Collaborative Practices**: Encouraging a culture of open innovation by adhering to FAIR principles, which promote the findability, accessibility, interoperability, and reusability of data.

By integrating these strategies, light source accelerators can maintain operational efficiency and reliability and support high-level scientific research while minimizing environmental impacts and improving sustainability. In practice,

[*] pardis.niknejadi@desy.de

these strategies are interdependent, often interacting to compound benefits.

FLASH1 UPGRADE STUDIES

At FLASH, the first high-gain FEL, an upgrade with a strong emphasis on the research and development towards a fully coherent light source is ongoing. The FLASH2020+ [6, 7] upgrade program pursue high-harmonic generation at high repetition rates in a superconducting seeded FEL. One of our main goals has been the study of advanced concepts in beam dynamics, accelerator, and FEL physics through realistic and reliable S2E simulations. Modeling externally seeded FEL requires noise suppression and smooth handshaking between multiple codes [8–12]. Due to the fine structure of the electron beam in the high harmonics seeding schemes, such as echo-enabled harmonic generation (EEHG) [13], the task involves simulating a large number of particles, which requires a large amount of memory and limits the possible studies to be conducted on electron beam properties, FEL performance, and properties of the output radiation in short time scale.

Scalable and Modular Software Solutions

In order to address the computational challenges posed by the FLASH1 upgrade, our team has developed a set of proprietary libraries designed specifically to manage the manipulation of beam, field, and lattice files in simulations. These libraries have undergone revisions based on feedback from our collaborators to prioritize modularity and extensibility, enabling their seamless integration into current projects while also ensuring adaptability for future requirements and compatibility with other simulation codes.

Additionally, we have created libraries dedicated to post-processing and analyzing simulation outputs. Furthermore, our analysis capabilities are enhanced by benchmarking processes that compare simulation results with available experimental data, establishing a robust framework for validating and refining our computational models based on empirical evidence. This approach has thus far effectively supported the ongoing development of the FLASH1 upgrade by optimizing simulation setups, such as grid size and precision, and facilitating the adjustment of computational load distribution, especially when using HPC resources to strike a balance between computational cost and accuracy. While this systematic investment in analytical and computational infrastructure may require additional work and resources at the outset, it pays off significantly.

Resource Management and Electron Beam with Fine Structure

Resource management is an essential strategy for sustainability in any field. It helps to enhance efficiency, reduce environmental impact, and control operational costs. Two main techniques are often used in beam dynamic simulations to reduce required resources: downsampling or macro particles and reducing particle files to essential slice parameters. The first involves reducing the number of particles tracked from billions to millions, significantly lowering computational overhead. Algorithms are employed to ensure that this reduction does not compromise the accuracy of the simulations. The latter, on the other hand, can reduce both input/output operations and computational load. This method is particularly effective when transitioning between simulation platforms such as Impact-Z [8], elegant [10], and Genesis 1.3 [11], as long as the longitudinal features of the beam match the slice width or simulation wavelength.

In conditions where high accuracy is required, such as EEHG simulations with high harmonic numbers (15, 30 and 75, for example), maintaining a detailed representation of the electron beam's fine structure is necessary to analyze bunching efficiency. In these cases, retaining the entire particle file is justified based on the objectives of the simulation, for instance, for simulations where the beam head and tail effect of an electron beam with varying energy profiles along the longitudinal axis (energy chirp) is being studied in detail as realistically as possible modeling is essential. At the same time, careful setup to ensure sufficient memory capacity is critical. Typical demand of a one-for-one Genesis 1.3 v4 simulation for EEHG setup with FLASH characteristic chirp (min and max) is highlighted in Fig. 1.

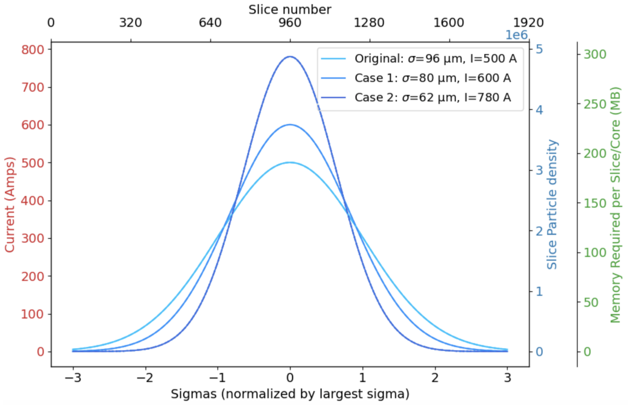

Figure 1: Current Distribution, Particle Density, and Memory Required for One-for-One Beam Slices in Genesis 1.3 v4. Memory requirements per slice are calculated based on the number of particles, each requiring 48 bytes for storage of 6D beam parameters with double precision . Case 1 illustrates the evolution of an electron beam with a 10 MeV/ps energy variation along its longitudinal axis, compressed during the EEHG process, from the original distribution. Case 2 shows a beam with 20 MeV/ps energy variation under similar conditions.

EEHG setup requires moderate peak current. Therefore, the current profile for such an experiment can be approximated as Gaussian distributions. As shown in Fig. 1, this leads to a significant load variation for slices/CPU cores. Nonetheless, the slice with peak current or maximum particle density indicates the simulation need. It defines the limit for the number of slices that can be assigned to one

CPU core, ensuring that the simulation runs successfully to completion. For the beam in Fig. 1, even the modulator section needs 500 to 1000 CPU cores. Depending on the target harmonic and the corresponding resolution in the radiator sections, the number of cores can be significantly increased, influenced by field memory requirements. Another solution would be utilizing high-memory cores, especially when the harmonic number and the number of grids are large, as shown in Fig. 2. A failed simulation in such a case could waste 1000s of CPU-hs. Effective resource management in such simulations, therefore, is feasible and cost-effective.

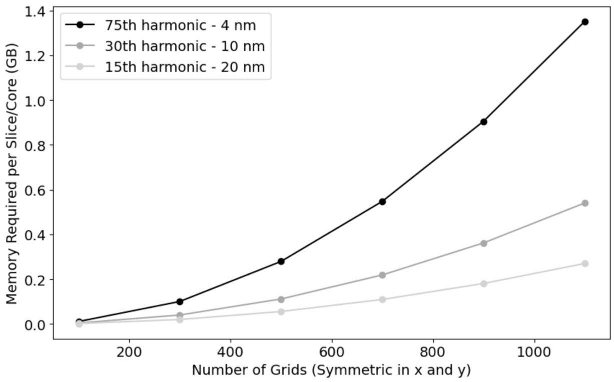

Figure 2: Memory Requirement in Radiators for Different Harmonics and Grid Sizes of FLASH1 EEHG simulations. Memory requirements per slice are calculated as the square of the number of grid times 16 bytes (real and imaginary field values with double precision).

Reusable Data and Data Driven Optimization

Realistic simulations, as discussed in previous sections, underscore the importance of a thorough understanding of the physics behind the problems being addressed, as well as the availability of computational infrastructure. Figure 3 illustrates the critical importance of detailed electron distribution data for assessing the FEL's sensitivity to parameter jitters, specifically timing in this example. Such studies or evaluations are indispensable for insightful upgrade planning and for supporting the diagnostics group. These would not be as effective with a simplified model.

As a next step, we plan to utilize the available virtual data for more extensive optimization using ML to develop robust data-driven decision-making strategies that enhance the operation of the upgraded FLASH1. Priority is given to simulations for which we have feedback from machine diagnostics with comparable accuracy. Moreover, the simulation data offers insight into online and real-time operational analysis demands. Given the multidisciplinary nature of this work, collaborations are essential and often resource-intensive. It is paramount that these projects be well-documented and made available publicly after an appropriate embargo period. For instance, simulation data from the FLASH2020+ project will be accessible here [14]. Furthermore, in collaboration with colleagues from high-energy physics, we evaluate our scientific carbon footprint [5]. Quantifying this value is important for decision-making processes and resource management.

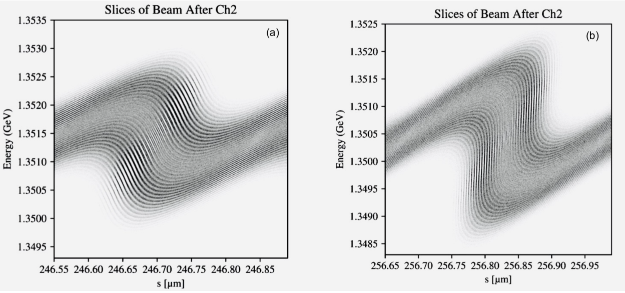

Figure 3: Impact of 30 fs (10 micron) Timing Jitter on Chirp-Induced Energy Spread and Bunching Efficiency: (a) illustrates the effects of a 10-micron delay between the seed laser and electron beam within an EEHG scheme. (b) when there is timing overlap. This figures presents simulations using a realistically modeled electron beam, highlighting how timing variations influence system performance in externally seeded free electron laser setups.

OUTLOOK AND SUMMARY

Through a detailed examination, this paper aims to initiate a collaborative dialogue to address some of the most pressing challenges in our field by leveraging recent trends in computational and AI technologies. While we've mainly focused on three of the six strategies, the remaining strategies are equally critical. For example, simulations requiring 1000 cores inherently necessitate using HPC facilities. The resource management examples for seeded FEL simulations in Genesis 1.3 v4 also apply to other complex fields, like modeling in Plasma Laser-driven Accelerator simulations involving the drive laser and the ultra-short witness electron beam with high particle density. A forthcoming task is to verify whether GPU optimization tools, typically favored for single precision, can also effectively enhance simulations where high accuracy is essential. This initiative represents another facet of our investment in custom software solutions. The strategic focuses discussed throughout this paper are designed to enhance the accuracy and reliability of our simulations and data analyses and boost the overall efficiency of research operations in a resource-aware manner.

ACKNOWLEDGEMENTS

We thank Enrico Allaria for engaging discussions on modeling and optimizing EEHG simulations. We also appreciate the support of the diverse team of the FLASH2020+ project at DESY. This work has been supported by BMBF within the project 05K19GUC. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS [15] at Jülich Supercomputing Centre (JSC).

REFERENCES

[1] "1st Large Language Models in Physics Symposium (LIPS)," DESY, hosted at DESY on Feb 21-23 2024 accessible via Indico. Accessed on: May 12, 2024.
https://ai.desy.de/events/lips
https://indico.desy.de/event/38849

[2] Google, "5 Principles for Cloud-Native Architecture," *Google Cloud Blog*, June 20 2019. Accessed: May 12 , 2024. [Online]. Available: https://cloud.google.com/blog/products/application-development/5-principles-for-cloud-native-architecture-what-it-is-and-how-to-master-it

[3] L. Lannelongue *et al.*, "GREENER principles for environmentally sustainable computational science", *Nat. Comput. Sci.*, vol. 3, pp. 514–521, 2023. https://www.nature.com/articles/s43588-023-00461-y

[4] V. Lang *et al.*, "Know your footprint – Evaluation of the professional carbon footprint for individual researchers in high energy physics and related fields,", 2024. doi:10.48550/arXiv.2403.03308.

[5] "Know your footprint (Kyf-2024)" survey[1], https://limesurvey.web.cern.ch/863499?lang=en

[6] L. Schaper *et al.*, "Flexible and Coherent Soft X-ray Pulses at High Repetition Rate: Current Research and Perspectives", *Appl. Sci.*, vol. 11, p. 9729, 2021. doi:10.3390/app11209729

[7] E. Ferrari *et al.*, "Status of the seeding upgrade for FLASH2020+ project", presented at the IPAC2024, Nashville, TN, USA. May 2024, MOPG15, this conference.

[8] J. Qiang *et al.*, "An Object-Oriented Parallel Particle-in-Cell Code for Beam Dynamics Simulation in Linear Accelerators", *J. Comput. Phys.* vol. 163, pp. 434–451, 2000. doi:10.1006/jcph.2000.6570

[9] T. Lang, "Chi2D and Chi3D", Accessed on: May 12, 2024. http://www.chi23d.com

[10] M. Borland, "ELEGANT: A flexible SDDS-compliant code for accelerator simulation", in *Proc. ICAP 2000*, Sep. 2000, Darmstadt, Germany. doi:10.2172/761286

[11] S. Reiche, "GENESIS 1.3: a fully 3D time-dependent FEL simulation code',, *Nucl. Instrum.*, vol. 429, pp. 243–248, 1999. doi:10.1016/S0168-9002(99)00114-X

[12] L. T. Campbell and B. W. J. McNeil, "Puffin : A three dimensional, unaveraged free electron laser simulation code", *Phys. Plasma*, vol. 19, 093119, 2012. doi:10.1063/1.4752743

[13] G. Stupakov, "Using the Beam-Echo Effect for Generation of Short-Wavelength Radiation", *Phys. Rev. Lett.*, vol. 102, p. 074801, 2009. doi:10.1103/PhysRevLett.102.074801

[14] DESY Public Data Repository, "DESY Public Data Repository." Available online: https://public-data.desy.de/datasets. Accessed on: May 12, 2024.

[15] D. Alvarez, "JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre", *Journal of large-scale research facilities*, vol. 7, p. A183, 2021. doi:10.17815/jlsrf-7-183

[1] The Kyf-2024 campaign link will be available until March 2025, offering scientists a tool to estimate their professional carbon footprint. It features options that allow for adaptability beyond the initial data. As more sustainability reports become available, they will be incorporated into future envisioned survey versions.