INTERACTIVE AI ON BWHPC – LESSONS LEARNED FROM BUILDING A LARGE-SCALE IMAGE ANALYSIS PLATFORM

ALEXANDER ZEILMANN[1,2]*, ERIK SCHNETTER[3] AND VINCENT HEUVELINE[1,2,3]

[1]INTERDISCIPLINARY CENTER FOR SCIENTIFIC COMPUTING, HEIDELBERG UNIVERSITY, HEIDELBERG, GERMANY

[2]DATA MINING AND UNCERTAINTY QUANTIFICATION GROUP, HEIDELBERG INSTITUTE FOR THEORETICAL STUDIES, HEIDELBERG, GERMANY

[3]UNIVERSITY COMPUTING CENTER, HEIDELBERG UNIVERSITY, HEIDELBERG, GERMANY

*ALEXANDER.ZEILMANN@IWR.UNI-HEIDELBERG.DE (CORRESPONDING AUTHOR)

ABSTRACT

AI and related technologies have rapidly reshaped the landscape of high-performance computing. Accordingly, many researchers in Baden-Württemberg face challenges due to the limited availability of interactive, GPU-ready, and service-oriented infrastructure within bwHPC. Drawing on systems we developed — KI-Morph and YoKI — we offer an experience-informed position on bwHPC's infrastructure remit, and the directions required to support modern research. While traditional, batch-oriented cluster infrastructure remains essential, we argue for a modular, interoperable approach that integrates clusters, cloud services, storage and more. We summarize today's landscape, introduce dimensions for diversification, and note that many of these capabilities already exist in mature industry platforms, underscoring feasibility and informing priorities within bwHPC. We note the risk that inaction will push researchers permanently to external providers and close with a call to collaborative enhancement of our infrastructure.

Keywords: AI, Interactive HPC, Interoperable Infrastructure, Cloud & Cluster Infrastructure Requirements

1. INTRODUCTION

Science has significantly changed in the recent years in all areas from scientific exploration, interdisciplinary collaboration and outreach activities. In particular AI-driven research increasingly depends on interactive, GPU-accelerated, and service-oriented workflows and applications that cannot be supported by traditional batch systems. We experienced these

challenges ourselves when developing KI-Morph[1], a platform for large-scale image analysis, and YoKI[2], the official LLM platform of Heidelberg University, as both applications require persistent availability, API access, workloads that come in bursts, and more. Accordingly, we want to share our perspective that draws on building and operating these systems as well as discussions with researchers and infrastructure teams. While the existing batch-oriented infrastructure remains a strength for many workloads and should be kept and developed further; nonetheless, new complementary capabilities are needed. We acknowledge that this is certainly not a simple task and will require significant work, time and funding. However, we believe that tackling these challenges is necessary and worth the effort as this will make sure that we further support modern research in Baden-Württemberg and beyond.

The remainder of this paper surveys existing infrastructure, introduces key dimensions for infrastructure diversification, motivates modularity for interoperability, discusses the threat of inaction, and concludes with a call to collaborative modernization.

1. CURRENT BWHPC INFRASTRUCTURE LANDSCAPE

Historically, almost all high-performance computing at our institutions was handled by SLURM-based batch computing. With new challenges and possibilities in the broader space of demanding computing, two major views of HPC have emerged. One view retains the historical meaning: SLURM-based, throughput-oriented batch computing on shared clusters, a perspective often held by technical experts. The other view broadens the term to encompass all highly demanding computing, including interactive, GPU-accelerated, and service-oriented workloads, a perspective often held by non-technical users. Neither perspective is incorrect; they reflect different vantage points. The historical view is solution-oriented, anchored in established tooling and operational models. The broader view is problem-oriented, centered on emerging user needs and application patterns. Depending on which view one adopts, opinions naturally diverge on the remit of bwHPC and whether the infrastructure we propose aligns with that remit. From our standpoint, the choice of definition and who operates the infrastructure matters less than ensuring that such infrastructure exists and is accessible. Because the proposed infrastructure addresses requirements that are pressing today, establishing it is more important than resolving terminological debates. Some

[1] https://ki-morph.de

[2] https://www.urz.uni-heidelberg.de/de/service-katalog/kuenstliche-intelligenz-ki/yoki

may judge the proposal a misfit for bwHPC when using the classic definition. However, even if bwHPC is not the perfect organizational home, it is currently our closest initiative to the infrastructure we advocate, which motivates presenting our work in this context.

2. CURRENT BWHPC INFRASTRUCTURE LANDSCAPE

We first review the existing infrastructure landscape within bwHPC to contextualize the requirements that follow. We briefly assess cluster, cloud, and storage offerings and their suitability for interactive, AI-centric workloads.

While bwHPC clusters are familiar to most readers, the bwVisu remote visualization service built on top of them may be new. bwVisu enables interactive applications to run on cluster nodes by streaming their user interface to researchers. Because access remains SLURM-based, this approach does not suit persistent, service-oriented applications that must be always-on and API-accessible. However, because tools like bwVisu are well supported, clusters often appear more mature than cloud offerings and are used for projects that would be better served by cloud-native services. KI-Morph is one such example: we built on the Helix cluster via bwVisu, because it was possible and cloud alternatives were not yet ready, not because the cluster was the ideal infrastructure.

Today, three cloud approaches relevant to our community exist but each is deficient for AI-centric, interactive workloads. bwCloud is intended as a shared cloud infrastructure for the bwHPC community, but it – to the best of our knowledge – currently does not contain GPU resources. Similarly, heiCLOUD, the cloud offering of Heidelberg University, does not yet provide GPU resources. The de.NBI Cloud[3] is a capable offering for the bioinformatics community in Germany, but its scope restriction prevents use for university-wide platforms such as YoKI or other AI services outside bioinformatics.

Across bwHPC, several storage solutions serve distinct purposes, including long-term archival, data exchange solutions, and storage of hot data for active computation. For hot data today, we effectively rely on a single class of solutions: the Large Scale Data Facility (LSDF) at KIT and the SDS Frontend at Heidelberg University. These solutions provide general-purpose storage and are well-suited for many research use cases. However, they can be ill-suited for very sensitive datasets, such as medical records, where specialized compliance and fine-grained access controls are required. They also do not easily support unauthenticated public file access, which some outreach projects require to distribute data openly.

[3] https://www.bw-cloud.org, https://heicloud.uni-heidelberg.de, https://cloud.denbi.de

3. DIMENSIONS FOR INFRASTRUCTURE DIVERSIFICATION

Rather than listing our own personal infrastructure requirements, we propose a broad problem-focused approach to enhancing the infrastructure. To this end, we believe that the requirements of users of the bwHPC infrastructure and equally, the demands researchers outside of it should be gathered and organized. Concrete technical choices should follow from this analysis and be prioritized by urgency and complexity. Against that backdrop, we suggest considering demands and solutions to be diversely distributed along a variety of dimensions.

One important dimension is the amount of compute required. We believe that this is handled well by the established classification of HPC clusters into tiers, where bwHPC clusters (Tier 2) fit well within the broader landscape. Another dimension is alignment to scientific fields. While our clusters are organized by domain today, it is debatable whether this yields meaningful technical advantages. With research increasingly interdisciplinary, infrastructure organized strictly by scientific field may be a poor fit for many projects. A particularly relevant dimension for modern applications is the accessibility–security trade-off. Accessibility and security are not opposites, but raising one often constrains the other in practice. Some domains, especially medical research, require stringent controls, while outreach projects demand maximum ease of access. Further, applications also vary along the time-model dimension: runtime and start-up characteristics. Some workloads run for days and can tolerate long start-up times, while other workloads require persistent services and millisecond-scale bursts, for which we currently lack a suitable, integrated solution.

Two further dimensions concern openness and ownership. Open-source versus closed-source should remain an area of real variability; excluding closed-source options a priori can unnecessarily limit viable solutions. Similarly, owning and operating hardware on-premises versus renting resources that may or may not be located in-house should be considered without any preconceived bias. Both dimensions should follow from the technical and organizational requirements above, not precede them.

Additional dimensions include developer enablement and operational complexity, I/O throughput and latency for data-intensive workloads, and service-level objectives such as uptime availability guarantees. Cost and funding models, data locality and egress patterns, and compliance constraints can also be considered design dimensions. Choices along all these dimensions should be made in dialogue with researchers, operators, and policy makers to align trade-offs with real needs. For dimensions that have a significant variability in the demands, the infrastructure should reflect this variability holistically. We deliberately avoid prescribing a specific technical solution, instead, we emphasize that these dimensions should serve as a framework for building a future-facing diversification of our infrastructure.

4. DESIGN PRINCIPLES FOR MODULAR INFRASTRUCTURE

Besides these dimensions we want to highlight the importance of modularity for the infrastructure. Modern research applications are composed of services that must interoperate across boundaries: data storage, compute, authentication, user-facing components and more. Creating isolated silos of infrastructure undermines reuse and raises operational burden in the long term. Instead, modularity should be a first-class design goal so that components can be composed for different projects and easily swapped as requirements evolve. Practically, this implies that clusters, cloud services and storage solutions are all able to be used together. The cloud can trigger jobs on the cluster, which can create files on the storage solution and that in turn can be accessed by the cloud. As the infrastructure evolves, and more components are added the modularity will become increasingly more important.

5. THE THREAT OF INACTION

It is tempting to adopt the narrow, traditional definition of HPC and continue investing solely in SLURM-based infrastructure. However, user requirements for a modernized infrastructure will not disappear by simply ignoring them. When researchers cannot satisfy these needs on bwHPC, they migrate to external providers that already offer suitable services. Once teams have established data pipelines, security reviews, and operational practices on external platforms, returning to bwHPC becomes costly and unlikely. Proactive investment in a more diverse, modular infrastructure is therefore essential to retain researchers and ensure long-term relevance.

6. CONCLUSION

The rise of AI and related technologies has rapidly expanded the community of researchers who need infrastructure tailored to interactive, data-intensive, and GPU-accelerated workflows. To meet these needs, we must diversify the infrastructure across clusters, cloud services, and storage, and make them modular and interoperable. Doing so will require sustained effort: structured interviews with researchers, careful evaluations, and open discussion between operators, policymakers, and domain experts. However, inaction is not an option; if needs are unmet, researchers will migrate to external platforms and may not return once their new workflows are established. We should embrace the changing HPC landscape as an opportunity to modernize, strengthen collaboration across institutions, and provide an infrastructure that enables cutting-edge, responsible research in Baden-Württemberg and beyond.