Evolution and Broadening of the National Analysis Facility at DESY

Christoph Beyer[1], , *Stefan* Bujack[1], , *Stefan* Dietrich[1], , *Thomas* Finnern[1], , *Martin* Flemming[1], , *Martin* Gasthuber[1], , *Sandro* Grizzo[2], , *Thomas* Hartmann[1], , *Johannes* Reppin[1], , *Yves* Kemp[1], , *Frank* Schluenzen[1], , *Christian* Sperl[1], , *Sven* Sternberger[1], , *Alexander*Trautsch[1], , *Christian* Voss[1], , and *Markus* Wengert[1],

[1]Deutsches Elektronen-Synchrotron DESY, Hamburg/Zeuthen, Germany
[2]Hochschule für Angewandte Wissenschaften, Hamburg, Germany

Abstract. The National analysis Facility (NAF) at DESY has constantly been evolving since its inception in 2007. Starting as a distributed computing platform between the DESY sites in Hamburg and Zeuthen, it has been serving the German HEP users as well as international collaborators since as a experiment-agnostic compute and data infrastructure. The technical implementations NAF have changed in a number of evolutionary steps over time to adapt to the changing requirements by its users as well as to the always changing technological landscape. While technological details have changed, central points to the NAF have been constant like data rather than plain compute being pivotal or like user support as a cornerstone. Followingly, we will describe the recent developments and updates in the NAF ecosystem. On the user side, further experiments have chosen the NAF as their computing platform and have build up their analyses pipelines ontop the NAF. On the operational side, effort has been made to further harden the security and increase monitoring and integration between compute and storage systems as complementary components of the NAF.

1 Introduction

1.1 The NAF: Scope and history

In 2007, the Helmholtz Alliance "Physics at the Terascale" was instantiated to bundle German activities in the field of high-energy collider physics. The alliance brought together 18 universities, two Helmholtz Centers, and one Max Planck Institute to work as a network of German research institutes working on LHC experiments, a future linear collider, and related phenomenology. Amongst a strong topical collaboration, also common infrastructure was created, the largest one was the National Analysis Facility (NAF).
The NAF complements the Grid by offering an complementary, orthogonal access to the data stored at DESY, on a compute setup optimized for analysis, fast turn-around and a good interactive experience. Since its inception, the data backend has been integrated within the WLCG allowing global ingress and egress of data via Grid workflows.
While the NAF was initially set up between the two DESY sites in Hamburg and Zeuthen, in 2013 a redesign took place that, amongst other, focused the setup to the Hamburg site only [1].

In 2024, the NAF comprises a batch system of around 250 servers, a dedicated project storage space in addition to NFS mounted Grid dCache storage [2], interactive login servers with ssh and FastX (browser based X11) access, and a Jupyter notebook setup [3].
The NAF has currently around 200 unique users per month. The initial user community stemming from ATLAS, CMS, ILC and LHCb has become larger: The global Belle II community has access to the NAF, as well as the legacy HERA experiments, and smaller DESY based axion experiments. Especially we will illustrate the usage of these more recent communities more in detail in this document.
From 2016 on, the technological base of the NAF compute cluster has been HTCondor [4]. We consolidated the NAF together with the Grid HTC computing cluster onto a common technological and administrative base. As such, both clusters only differ by detail cluster configurations. A further operational unification into a single logical cluster had been considered, but which had been deferred due to the different load profiles. Instead, we have chosen to leapfrog and embed the clusters into an Interdisciplinary Data and Analysis Facility (IDAF), which will offer users a seamless access to the compute and storage infrastructures with compute and data being brokered.

1.2 The NAF and the IDAF at DESY

The IDAF is at its core an integrated storage and compute system, where both components are complementing each other.
With respect to computing, the IDAF consists currently of three compute clusters serving different user needs but which are about to be integrated into a common tool set for the user communities at DESY - while experienced users will still be able to address specific clusters directly.
Storage is provided by three major storage systems as well. For one, a GPFS/Spectrum Scale instance is providing fast scratch space. Long-term storage including tape archival is handled by a number of dCache instances [2]. Finally, a AFS/YFS cluster still serves user homes in parts of the IDAF.

1.2.1 Compute Infrastructure

Since 2003, DESY operates the Grid cluster, which serves the production needs of LHC experiments ATLAS, CMS and LHCb, as well as a raw data center for Belle II and ILC.
In 2007, the NAF was set up as complement to the Grid cluster to allow users a faster and more interactive analyses workflow and job turnaround with results becoming available in hours rather than days on the grid. While the NAF and Grid are still somewhat HEP centric and large scale data acquisition and analysis at DESY had initially been the domain of high energy physics, this has changed with the PETRA III facility and the European XFEL becoming operational in the early 2010s. These photon facilities are equipped with new high frequency and high resolution detectors and have been since dominating the on-site data and computing requirements [5].
To serve the on-site photon experiments and users, DESY IT started in 2011 the Maxwell HPC cluster as a platform for massive parallel simulations and analysis jobs with high bandwidth, compute or memory needs. The Maxwell HPC cluster was integrated into the DAQ of photon science experiments similar to the large scale compute clusters near the HEP experiments at their respective T0s. The Maxwell HPC grew in size as grew the data of the photon science experiments and is now comprised of over 900 servers, of which around 200 servers equipped with 400 GPU. All systems have InfiniBand interconnect for low latency inter process communication and fast access to an IBM GPFS file system. The cluster is

managed using SLURM, and has similar access methods as the NAF.

1.2.2 Storage Infrastructure

Complementing the different compute systems, NAF, Maxwell HPC, and Grid, different storage systems are the second constituting core of the Interdisciplinary Data and Analysis Facility (IDAF).

Intended as fast scratch space, GPFS/SpectrumScale storage clusters are available as shared file systems on the user facing clusters of the NAF HTC and Maxwell HPC. A previously operated BeeGeeFS storage instance in Maxwell has been subsumed by a general GPFS storage.

As long-term storage, DESY IT operates a number of dCache instances, which are also available as shared file systems on the compute clusters and user facing servers. Additionally, world-wide data exchange is facilitated through grid-workflows on dCache instances allowing for easy data ingress and egress. Crucially for on-site experiments and groups but also becoming more relevant for HEP experiments is the cost-effective tape archival for data preservation provided by the dCache instances.

Furthermore, DESY operates a AFS/YFS cell with the storage being intended for easy small-scale global data exchange. Similarly, DESY IT operates a user file hosting and exchange service based on Nextcloud with dCache providing the backend storage.

While until mid 2010s HEP data usage dominated at the IDAF, with the EuXFEL becoming operational photon science has become by far the dominating scientific community with about 150PB of data under dCache management.

In general, data are available to users by POSIX means, i.e., logical addressing via mounted paths and file I/O by corresponding kernel system calls. Due to the wide range of user groups, experiences and especially tools, POSIX-like file access is the common foundation supported by all. Also the I/O system calls and network protocol iterations, optimized oevr decades, ensure high performance in data read and write operations. File access control is handled by user and group IDs as well as, optionally, extended ACLs. This approach allows the IDAF to be open and being interoperable for the wide and diverse set of users and communities and being interdisciplinary in its core.

In the past we had differing mount namespaces on our NAF HTC cluster and our Maxwell HPC cluster due to their parallel growth. In the preparation of the IDAF and update of the Maxwell storage backend, we have consolidated these namespaces into a common set of paths. Thus on all user facing servers the same mount namespace is available, so that users can easily switch between clusters encountering the same paths.

1.2.3 User Access

While compute and data are intrinsically local, NAF and IDAF serve the German as well as international scientific communities. Initially, access and user management to the NAF had been managed via X.509 user certificates with users being authenticated and managed by their personal certificates.

Each user gets a local user identify assigned as well as a primary group within the DESY user namespace, optionally additional secondary groups are possible. With the general momentum away from X.509 certificates towards more flexible access delegation protocols, user authentication and authorization for NAF and IDAF access is being integrated with the established community federated Single-Sign-On (SSO) and Identity-Access-Management

(IAM) frameworks.
Both authentication & authorization approaches, legacy X.509 and token based, are technical open to various world-wide federated scientific instances like Helmholz Cloud, EGI Check-In and similar. After validating a new user request, a local user identity is created and mapped to the corresponding IAM identity. Followingly, file and compute process ownership by the user is handled under the user's UID and GIDs, thus allowing a seamless use of tools developed with POSIX semantics and I/O APIs in mind.

1.3 General Concepts

By adhering to intrinsic local user, data and process namespaces, we avoid the complications of attempting to construct consistent global namespaces. As federated user identifies are flexible mapped to the local user namespace, the NAF can be tightly integrated into the international computing infrastructure allowing roaming users to easily flock to NAF resources.
As the scientific communities in the IDAF, HEP and photon, are data centric with embarrassingly parallel or I/O heavy workloads, the IDAF is oriented with data in mind. Consequently, the storage infrastructure is aligned in a quality-of-service (QoS) hierarchy with respect to speed as well as cost effectiveness.
As computing utilization and efficiency suffers from data I/O latency, we aim to keep data and compute resources as close to each other. Data can easily be staged onto dCache storage instances, let it be from remote sources via Grid workflows or from local tape archival. Intermediate data from reprocessing or simulation is then to be stored on the fast GPFS scratch space as high-performance level in the QoS storage hierarchy.

2 Operations

2.1 Compute Cluster

Both IDAF HTC clusters are based on HTCondor [4]. Both clusters share the same hardware base (with slight variations in some generations) and technical configurations except for specific HTCondor details. Originally, we had intended to subsume both use, Grid and NAF, in a common HTCondor cluster, but we decided to keep both clusters separate due to heterodox user applications and I/O approaches. In the meantime, work is ongoing for back-filling NAF compute resources during low utilization with opportunistic Grid jobs.

2.1.1 NAF HTCondor Set Up

In contrast to most HTCondor clusters, we use in the NAF a remote submit approach to separate user instances from the HTCondor access points (AP). This allows us to more easy switch HTCondor access points for load management or maintenance without interrupting users on their login nodes. We operate for groups login nodes a number of these workgroup servers to which users can connect via ssh login. In addition we operate dedicated hardware login nodes with GPUs or larger memory for some of our groups.
To allow for a faster job allocation to execution nodes, we operate multiple collector instances for different use cases.
Since we also run interactive work loads on the NAF, we have set up a fast track collector, which brokers time critical jobs like Jupyter notebook jobs to dedicated job slots. Thus,

we can reach near interactive start of notebook jobs in parallel to non-time critical batch jobs. Execution nodes come in different flavours, which can be selected by the users during submission. The vast majority of the execution nodes are compute optimized running without SMT, with a core to memory ratio of at least 1:4GB. These default jobs limited to three hours run time can overflow to NAF shares nominally belonging to other groups. Larger jobs are bound to the user's group quota on the NAF resources.

2.2 Monitoring

The NAF's tightly integrated compute and storage cluster make a integrated monitoring necessary. We tightly monitor the dCache storage clusters' transactions by aggregating all storage events [2]. These storage evens are collected in a Elastic Search database and are used in parallel for alarming as well as issue tracing with Kibana and Apache Spark analyses. Similarly, for IDAF HTCondor clusters we collect HTCondor job events. Like for the storage instances, we use these job status transitions or resource usage information for alarming and issue debugging (See an example in Figure 1).

To integrate both cluster event information into a common monitoring views, we implemented an eBPF-based toolkit to collect such file access metrics on execution points. File access and throughput information are collected on a user and PID level at the kernel and processed in user space before sending the metrics to the monitoring infrastructure. A schematic view of the metric flow is shown in Figure 2a.

With these very fine grained metrics, we can trace user job I/O distributed over the compute cluster and intermesh these with the corresponding storage pool metrics. An example in Figure 2b shows the number of file reads and writes operations by one user on a compute node with local writes of cache files and remote reads of input data.

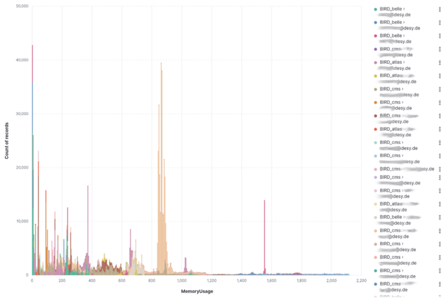
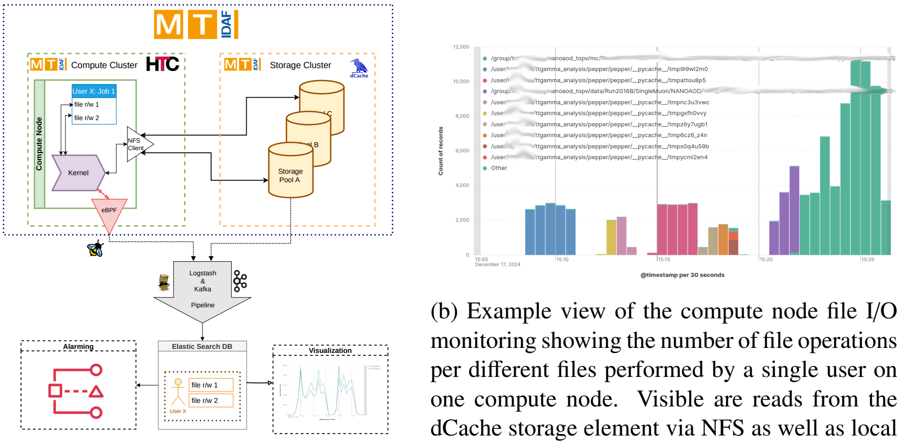


Figure 1: Example for HTCondor job events. Visualization of maximum memory RSS values for jobs on the NF HTCondor cluster over three days. The breakdown is by group and by user name.

2.3 User Support

As the IDAF's user base is wide ranging in both the experience level as well as with respect to scientific background, user support and education is a fundamental need for operating the IDAF. Above the user level, we regularly meet with the NAF stakeholders, i.e., computing experts of the various physics groups, to exchange updates and discuss problems.

Complementing day to day support, we are engaged in user education and workshops like user workshops on sustainable computing [11].


Compute Cluster
Storage Cluster
dCache
Compute Node
User X: Job 1
file r/w 1
file r/w 2
NFS Client
Kernel
eBPF
Storage Pool A
Pool B
C
Logstash & Kafka
Pipeline
Alarming
Elastic Search DB
file r/w 1
file r/w 2
User X
Visualization
12,000
10,000
8,000
6,000
4,000
2,000
0
Count of records
/group/[redacted]/nanoaod_topv/mc/[redacted]
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmp9l9wl2m0
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmpattou8p5
/group/[redacted]/nanoaod_topv/data/Run2016B/SingleMuon/NANOAOD/[redacted]
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmpnc3u3vwc
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmpgxfh0vvy
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmpz6y7ugb1
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmp6cz6_z4n
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmps0q4u59b
/user/[redacted]/ttgamma_analysis/pepper/pepper/__pycache__/tmpycmi2en4
Other
15:05
December 17, 2024
15:10
15:15
15:20
15:25
@timestamp per 30 seconds

(a) Schematic view on the metric flow for fine grained compute I/O monitoring.

(b) Example view of the compute node file I/O monitoring showing the number of file operations per different files performed by a single user on one compute node. Visible are reads from the dCache storage element via NFS as well as local writes of cache files to the local disk.

Figure 2: NAF compute cluster monitoring for a consolidated view in combination with the storage pool monitoring

3 AF Applications

In addition to native parallel batch processing of data, a number of other use cases have emerged over the last years ontop the NAF compute and storage clusters A crucial point has been, that the NAF in its fundamentals is not designed for a specific use case or with a dedicated user group in mind, but as open, flexible and interdisciplinary as possible while reducing complexity, so that each of our groups and users can capitalize the resources for their specific needs.

3.1 Jupyter Notebooks

As Jupyter notebooks have become a favourite tool for interactive analyses, we have deployed a scalable Jupyterhub solution [3] as additional interface to the NAF. As Jupyter notebooks are run in normal HTCondor jobs, all mount namespaces and resources are available like for classic batch jobs.

3.2 Containers as application deployment

We see containers as tools of choice to separate application OS requirements from underlying OS requirements of the server installations. Similarly, the larger HEP groups have all abstracted their applications with containers to be flexible with respect to the target platforms. While larger groups and communities have established workflows and experiences in using containers, smaller groups often lack the manpower for own elaborated set ups. Thus, we teamed up with our smaller user groups to develop shared workflows for building and distributing containers via Gitlab pipelines with CVMFS [6] and scratch space as target file systems.

3.3 Columnar Analyses

Columnar analyses are currently driven by HEP groups as tool for a fast turnaround with an optional interactive experience in scientific data processing. Here, we have observed issues with problematic file access patterns and are now engaged with developers on the user side to understand their I/O approaches.

4 Axion experiments and their usage of the NAF

As mentioned before, smaller groups with experiments on-site have become major users of NAF resources over the last years. These experiments like ALPS II [7], MADMAX [8] or IAXO [9] focus on axion searches as well as non-perturbative QED studies like LUXE [10]. They take advantage of the flexible design of the NAF and profit from the close collaboration with DESY IT. Instead of an experiment specific approach, we aim to be interdisciplinary also with respect to the experiment groups, thus pooling computing and storage needs and offering mutual usable solutions to common requirements. The collaboration was initiated between ALPS II and DESY IT and extended to MADMAX and IAXO over time.

4.1 Experiment Pipelines

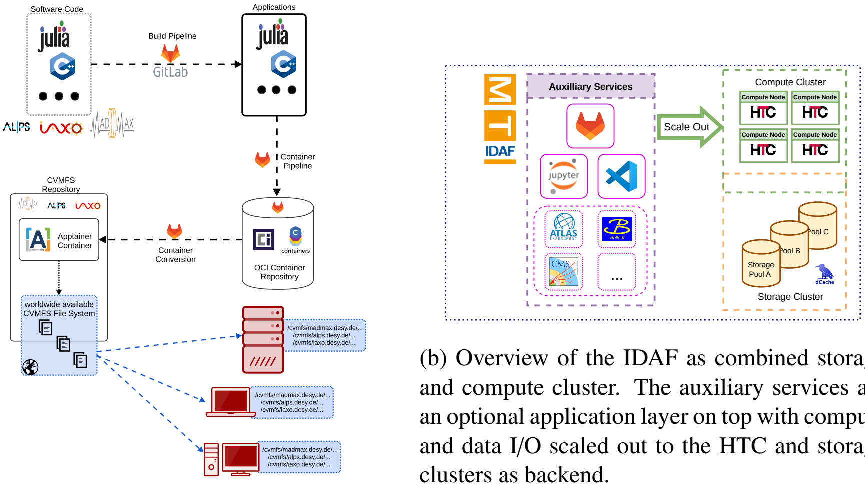
To optimize manpower and resources, DESY IT and ALPS II scientists have been tightly collaborating on computing issues. Together, we set up scalable software and container pipelines as sketched in Figure 3a and which has been serving as blueprint for other experiments.
Besides reducing the workload for us by using shared approaches, we benefit from the close collaboration with the on-site experiments with tight feedback loops like active feedback during pre-production phases, where MADMAX members identified a crucial kernel bug.

5 Outlook

Due to its interdisciplinarity, the NAF differs from emerging experiment specific analysis facilities. While such silo analysis facilities are designed to serve specifically an experiment's established analysis models, we strive to avoid lock-ins to particular set ups. However, as users' analysis models become more complex going beyond plain batch jobs, we plan to integrate such evolving models into the NAF by auxiliary services.
Here, we plan to offer experienced user groups an interface to deploy their helper applications and integrate these with the NAF compute cluster for scale out.

5.1 Auxiliary Services

As sketched in Figure 3b we see such auxiliary services as an optional application layer on top the compute and storage cluster. The Jupyterhub and Jupyter notebook setup can be seen as a blueprint for other possible auxiliary services as they are not doing any compute heavy work themselves as such but scale out to the compute cluster. Separating auxiliary in a third pillar of the IDAF, additional access and usage models can be flexible deployed. With DESY IT managing the scale out to the NAF compute cluster as interface, deploying such auxiliary services will not be limited to DESY IT managed generic services but potentially also can be utilized by user groups for their specific needs ontop the NAF.
As Kubernetes has becoming an established application platform, we are working on a dedicated Kubernetes auxiliary cluster available to dedicated users. Since Kubernetes is less



(b) Overview of the IDAF as combined storage and compute cluster. The auxiliary services are an optional application layer on top with compute and data I/O scaled out to the HTC and storage clusters as backend.

(a) ALPS-II Gitlab pipeline for software deployments to CVMFS.

Figure 3: NAF as infrastructure of three inetrlocking components: data storage, bulk compute scale out and lightweight auxiliary services.

suited to compute and data heavy scale out, when ensuring proper user and application isolation, it is intended for auxiliary services only with scale out to the batch system. Also due to security isolation constraints, POSIX mounts can naturally not be made available but application would need to use user space protocol realizations for storage access.
As the use of Kubernetes including preparing complex container set ups can be challenging, we do not see ordinary users analyses as prime target but rather computing experts in our scientific groups. These users will be able to deploy their applications onto the auxiliary cluster, which then can be used by their analyst colleagues, e.g., serialization services for transforming ROOT ntuples into columnar formats.

6 Summary

The National Analysis Facility at DESY has evolved over its nearly 20 years of existing into an interdisciplinary platform for data centric computing. Under the umbrella of the Interdisciplinary Data and Analysis Facility, the NAF is nowadays a tool used by a wide range of users from a large number of scientific groups and users.
Due to the data centric nature, storage and compute are equally important parts and require a combined and integrated monitoring and alarming.
As the NAF, under the umbrella of the IDAF, is designed to be flexible and experiment agnostic, additional scientific user groups have settled on the NAF as their analysis platform profiting from the mature and agile set up adaptable to their needs. We aim to streamline these needs in collaboration with our user groups to reduce workloads for DESY IT as operator and our scientific groups as users of the NAF.
Over the years a third pillar, in addition to the compute and storage components, emerged

in the form of auxiliary services. These auxiliary services provide additional functionality to the NAF and scale out their actual compute and data workloads to the established clusters. With this resilient approach, we can optimize and evolve the NAF within the IDAF where the functionality of data centric computing is at the core but combined with the flexibility to adapt to the ever changing landscape in scientific computing.

References

[1] A. Haupt, Y. Kemp and F. Nowak, J. Phys. Conf. Ser. **513** (2014), 032072 https://doi.org/10.1088/1742-6596/513/3/032072

[2] T. Mkrtchyan, K. Chitrapu, V. Garonne, D. Litvintsev, S. Meyer, P. Millar1, L. Morschel, A. Rossi and M. Sahakyan EPJ Web Conf., **251** (2021), 02010 https://doi.org/10.1051/epjconf/202125102010

[3] Reppin, J., Beyer, C., Hartmann, T. *et al. Comput Softw Big Sci **5** (2021), 16*

*[4] D. Thain, T. Tannenbaum and M. Livny Concurrency - Practice and Experience **17** 2005, 323-356*

*[5] C. Beyer, S. Bujack, S. Dietrich, T. Finnern, M. Flemming, P. Fuhrmann, M. Gasthuber, A. Gellrich, V. Guelzow and T. Hartmann, et al. EPJ Web Conf. **245** (2020), 07036 https://doi.org/10.1051/epjconf/202024507036*

*[6] L Promberger, J. Blomer, V. Völkl & M. Harvey CernVM-FS at Extreme Scales EPJ Web of Conf. **295** 2024*

*[7] A.D. Spector et al. ALPS II technical overview and status report **2017** https://doi.org/10.3204/DESY-PROC-2009-03/Spector_Aaron*

*[8] B. Majorovits, J. Redondo et al. MADMAX: A new Dark Matter Axion Search using a dielectric Haloscope **2017** https://doi.org/10.3204/DESY-PROC-2009-03/Majorovits_Bela*

*[9] E. Armengaud et al. Conceptual design of the International Axion Observatory (IAXO) Journal of Instrumentation **9** May 2014 https://doi.org/0.1088/1748-0221/9/05/T05002*

[10] L. Helary et al. LUXE: A new experiment to study non-perturbative QED 2023 https://doi.org/10.1393/ncc/i2023-23024-y
*CernVM-FS at Extreme Scales EPJ Web of Conf. **295** 2024*

*[11] J. Alimena et al., Sustainable computing workshops in high-energy physics at DESY Front.Comput.Sci. **6** (2024) 1502784 https://doi.org/10.3389/fcomp.2024.1502784*