Report from Dagstuhl Seminar 24462

Research Infrastructures and Tools for Collaborative Networked Systems Research

Georg Carle^{*1}, Serge Fdida^{*2}, Kate Keahey^{*3}, Henning Schulzrinne^{*4}, and Sebastian Gallenmüller^{†5}

- 1 TU München Garching, DE. carle@tum.de
- $2 \hspace{0.5cm} Sorbonne \hspace{0.5cm} University Paris, \hspace{0.5cm} FR. \hspace{0.5cm} {\tt serge.fdida@sorbonne-universite.fr}$
- 3 Argonne National Laboratory, US. keahey@uchicago.edu
- 4 Columbia University New York, US. hgs@cs.columbia.edu
- 5 TU München Garching, DE. sebastian.gallenmueller@tum.de

— Abstract -

This report presents the program and outcomes of Dagstuhl Seminar "Research Infrastructures and Tools for Collaborative Networked Systems Research" (24462). The seminar brought together experts from the network and distributed systems testbed community, scientists who rely on testbeds for their research, and representatives from funding agencies. It focused on bridging the gap between the services provided by large-scale testbed infrastructures and the needs of researchers conducting cutting-edge experiments. Discussions centered on enhancing the value and impact of research infrastructures by improving collaboration, streamlining experiment workflows, and developing testbed-agnostic tools. The goal was to make experimental research more modular, adaptable, and reproducible, ensuring that experiments and evaluation software can be easily modified, extended, and ported across different testbed environments. Key topics included strategies to improve research quality, reproducibility, and reusability, enhance the discovery process, and maximize the efficient use of research infrastructure resources.

Seminar November 10–13, 2024 – https://www.dagstuhl.de/24462

2012 ACM Subject Classification Networks \rightarrow Network performance evaluation

Keywords and phrases Research Infrastructures, Testbeds, Reproducibility, FAIR: Findability, Accessibility, Interoperability, and Reuse of digital assets, Infrastructure usage and sharing,

Artifact Evaluation, Optimizing reuse of data

Digital Object Identifier 10.4230/DagRep.14.11.60

1 Executive Summary

Georg Carle (TU München – Garching, DE) Serge Fdida (Sorbonne University – Paris, FR) Kate Keahey (Argonne National Laboratory, US) Henning Schulzrinne (Columbia University – New York, US)

Research infrastructures should evolve towards advanced scientific instruments that offer a vital insight to the underlying information in improvising the understanding of science methodologies and practices as they reliably and precisely help the scientist to measure the subject of their investigations. The Dagstuhl Seminar participants strongly agreed

Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

REPORTS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

^{*} Editor / Organizer

[†] Editorial Assistant / Collector

Research Infrastructures and Tools for Collaborative Networked Systems Research, *Dagstuhl Reports*, Vol. 14, Issue 11, pp. 60–91

Editors: Georg Carle, Serge Fdida, Kate Keahey, and Henning Schulzrinne

DAGSTUHL Dagstuhl Reports

that large-scale research infrastructures are essential for providing scientists with access to specialized, advanced resources enabling cutting-edge experiments. As a result, the following key conclusions were drawn:

- A. Strategic Investment & Community Engagement: Research infrastructures represent a vital and long-term investment that demands active participation from research communities, sustained human capital development, and financial sustainability.
- B. **Open Access & Data Sharing:** While open access to shared physical infrastructure is essential, access to open research data is equally critical. Digital sharing of scientific results accelerates innovation, enhances reproducibility, and strengthens FAIR (Findable, Accessible, Interoperable, and Reusable) data sharing through metaservices.
- C. Amplified Impact & Network Effects: Research infrastructures inherently complement and amplify each other, creating a synergistic network effect. This interconnectedness fosters a more rigorous scientific approach and methodology, driving greater collaboration and knowledge advancement.

The results of the seminar include the following key recommendations:

Key Recommendations

- 1. **Define clear scientific objectives:** Research infrastructures must explicitly articulate their scientific goals and establish a well-defined set of research questions to address.
- 2. Foster a strong scientific community: The success of research infrastructures depends on strong community engagement. Support measures are essential to strengthen and sustain the scientific community. The effort of their support teams should be better recognized.
- 3. **Implement EasyFAIR principles:** Adopting an EasyFAIR framework offering comprehensive and automated support for researchers is crucial to ensuring and leveraging the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) of digital assets. Additionally, open research data and reproducibility should be mandated by funding agencies and scientific societies. Scientists making an effort to share their research data should be rewarded.
- 4. Enhance reproducibility: Reproducibility is a critical priority. Concrete methodologies must be established to ensure comparability of experimental results across different research infrastructures.
- 5. **Multi-year investment strategy:** Research infrastructures should be properly articulated, designed and supported according to a longer-term roadmap and a sustained investment strategy.
- 6. Establish common abstractions: Standardized models should be widely adopted for describing experiments and associated frameworks, including information models, data models, and ontologies.
- 7. **Improve findability and accessibility:** The discovery and accessibility of research infrastructures and testbeds should be enhanced through comprehensive catalogs detailing available hardware and functionalities. It is also essential to assess how new and planned infrastructures contribute to scientific diversity.
- 8. **Define standardized evaluation criteria:** A clear and well-defined set of evaluation criteria is necessary to assess the relevance and impact of research infrastructures. The outcomes of the "Testbed Evaluation" World Cafe (cf. Section 7) provide a valuable basis for these criteria. Standardized assessment frameworks should be established for different categories of testbeds.

- 9. **Optimize user experience:** Usability for researchers must be a priority. An innovative metric, *Time to First Experiment (TTFE)*, can be used to measure the efficiency of infrastructures in enabling rapid experimentation and adoption. Education and training should be a strong component of research infrastructures.
- 10. Ensure interoperability and openness: Strong support for interoperability between testbeds is crucial as well as using open components as often as possible, as it supports the ability to easily port experiments across different infrastructures.
- 11. **Promote flexibility and adaptability:** Provisioning of ready-to-use experimental platforms: Instead of providing merely the fundamental resources of an experiment, the testbed should provide a experimental templates that researchers can use to perform their own research. This experimental template or *"blueprint"* can be used by researchers to answer specific research questions. These blueprints are not static but *melleable*, i.e., researchers are encouraged to adapt and extend them to fit their needs. The concept of *malleability* includes facilitating the modification of software artifacts and support of composability.
- 12. Support sustainable development goals (SDGs): Large-scale research infrastructures contribute directly to the SDGs by optimizing the efficiency of hardware resource usage and improving workflows from experiment design to result dissemination. Additionally, insights gained from research findings can enhance global technical infrastructure, further supporting SDG objectives.

2 Table of Contents

Executive Summary	
Georg Carle, Serge Fdida, Kate Keahey, and Henning Schulzrinne	60
Aims and Scope	65
Organisation of the Seminar	67
Overview of Talks	67
SLICES, Research Infrastructure Defined as a Scientific Instrument Serge Fdida	67
Chameleon: New and Noteworthy Kate Keahey	68
Orchestration for Reproducibility and Metadata Management Sebastian Gallenmüller	68
Research Infrastructure Opportunities at NSF $D_{\text{res}} = M_{\text{rel}} h_{\text{rel}}$	60
MERIF Mid-scale Experimental Reserach Infrastructure Forum	09
Paul Michael Ruth	69
6G Visions and Possibilities for EU-US Collaboration on Network and Internet Technologies Jorge Gasos	70
Data Management Tools, Workflows, and Recommendations from NFDIxCS Michael Goedicke	70
FABRIC Paul Michael Ruth	71
Post-5G Experiments <i>Damien Saucez</i>	72
Joiner and NDFF Andrew W. Moore	73
SPHERE Research Infrastructure Jelena Mirkovic	74
European ESFRI Roadmap and Research Infrastructure Landscape (Digit Group) Hakima Chaouchi	75
Breakout Session	75
Working Group: Reproducibility Jelena Mirkovic and Kate Keahey	75
Working Group: Workflows and Software in Research Infrastructures Damien Saucez	76
Working Group: Artificial Intelligence and Machine Learning Digital Twins Georg Carle and Walter Willinger	78

Working Group: Testbeds and Science Hakima Chaouchi and Walid Dabbous	79
Working Group: Improving Testbed User Experience Terry Benzel and Paul Michael Ruth	81
World Cafe	82
World Cafe Theme: 5G/6G and 3GPPJörg Widmer	82
World Cafe Theme: Testbed Evaluation Georg Carle	83
World Cafe Theme: Long-running Experiments Sebastian Gallenmüller	84
World Cafe Theme: Better TestbedsBjörn Scheuermann	86
World Cafe Theme: SustainabilityTom Barbette	87
World Cafe Theme: International CollaborationPaul Michael Ruth	89
Plenum discussions	89
Participants	91

3 Aims and Scope

Experimental research on networked systems requires a suitable research infrastructure to perform experiments. For many years, the dominating approach of research groups was to create a specific experimental setup tailored to the needs of specific experiments, e.g., in the context of a PhD thesis. The community is well aware of the obvious shortcomings of this approach. One shortcoming is that while the time and effort to set up the needed infrastructure is high, the fact that the setups are created independently not only means unnecessary duplication of work, but also heterogeneity, with details frequently not documented in publications, that may lead to difficulties reproducing experiments of other scientists. The need to address such shortcomings has been identified for a long time, e.g. in the 2003 workshop of ACM SIGCOMM "Models, Methods and Tools for Reproducible Network Research" [1]. To overcome such challenges, the networked systems community built large testbed research infrastructures. In the US, this includes testbeds such as GENI [2], FABRIC [3], Chameleon [4], and CloudLab [6], and in Europe, this includes initiatives such as Fed4Fire [8] and SLICES [7]. While these initiatives demonstrated significant progress, a relevant gap seems to exist between what the testbed research infrastructure community provides as services, and what scientists involved in cutting-edge experimental research actually need, and that closing this gap would make these testbeds the natural research infrastructure, where the leading experimental research is performed.

The existing large-scale testbed research infrastructures provide highly valuable resources for scientists, accompanied by solid tools for resource allocation. However, the basic services offered by the testbeds cannot fully cover the requirements to orchestrate complex experiments needed by researchers to be most productive. One area of possible improvement is experiment control. So far, scientists who use a large testbed for specific experiments need to solve how to orchestrate their experiments, how to collect, process, and store the data produced by the experiment, and how to add metadata to support other scientists. Consequently, there is a high heterogeneity concerning the artifacts of specific experiments. Typically, an artifact evaluation committee (AEC) evaluates the artifacts (software and data with metadata) for papers that provide them. The heterogeneity and lack of generally available tools for the experiment workflow lead to a high effort: (1) for those who prepare the artifacts, (2) for those who review the artifacts (as part of the AEC work), and (3) for those who want to make use of the artifacts, e.g., when comparing own artifacts with previously published artifacts. This approach clearly does not scale, in particular in the context of an increased data-driven science powered by AI/ML.

The seminar brought together the following scientific communities:

- The network testbed community, that builds testbed research infrastructures, operated as a service for other scientists.
- Scientists that need such research infrastructures to perform their experiments with domain-specific components: computer networks, cloud and edge infrastructure, highperformance computing systems, operating systems, including high-performance I/O.
- The tools community, that develops tools for performing experiments and evaluating the data generated by these experiments.

The seminar aimed to identify an approach that combines measures suitable to overcome the identified problems.

A possibility to improve the usability of research infrastructures is the creation of *blueprints*, that provide well-structured templates for scientific experiments involving all phases of experiment design, execution, data and metadata generation, processing, evaluation,

and publication. The experiment workflows use a structure designed for efficient reuse with enough flexibility to adapt to new experiment requirements by changing specific parts or tools used in the workflow – we call this property *malleable* experiment workflows By introducing a research infrastructure with a framework and associated tools that support experiment workflows that can be modified, extended, combined, compared, and ported to other testbed environments, a number of benefits can be achieved:

- Scientists do not need to create their own tooling and workflows, they instead can rely on an existing, extensible workflows and can concentrate their effort on the specific problems they want to investigate.
- Scientists that set up experiments in a suitable testbed benefit from the additional tools, and can submit artifacts for evaluation by making the artifacts ready for execution by an evaluation committee in a suitable testbed.
- For members of an artifact evaluation committee, the review task becomes more efficient when artifacts to be evaluated come with experiments controlled by known tools.
- If published papers with associated experiment artifacts use such a framework and tools, other scientists can more easily build on existing work, by creating forks of the original experiment, comparing it with modified versions, or combining it with other artifacts.
- Experiment software, data, and metadata can be more easily found, understood, and combined with tools installed in the testbed that automate, in a comprehensive manner, the needed steps of the scientific work.

There exists a significant amount of experiment artifacts that can be executed in specific testbeds. For example, the Chameleon testbed [4] provides Trovi [5], a repository for experiments. However, as the testbed provides tools for resource allocation, but not for experiment control, these experiments do not have a common experiment control structure, as no software framework for experiment control is provided by the testbed. This makes it difficult for other scientists to do follow-up work, for example, performing experiments that combine components of two existing experiments. Concerning frameworks for experiment control was published at CoNEXT'21 [9]. This framework and toolset for experiment control was published at Technical University of Munich, meeting the needs of its many local users. Concerning Artifact Evaluation, the biennial ACM Symposium on Operating Systems Principles SOSP started a collaboration with the Cloudlab testbed [10].

References

- 1 Workshop on Models, Methods and Tools for Reproducible Network Research (MoMeTools), Program Co-Chairs: Georg Carle, Klaus Wehrle, Hartmut Ritter, Karlsruhe, Germany, August 25, 2003, In conjunction with ACM SIGCOMM 2003, https://conferences. sigcomm.org/sigcomm/2003/workshop/mometools/.
- 2 GENI. https://www.geni.net/. Last accessed 2025-02-20.
- 3 FABRIC. https://fabric-testbed.net/. Last accessed 2025-02-20.
- 4 Chameleon. https://www.chameleoncloud.org/. Last accessed 2025-02-20.
- 5 Trovi. https://www.chameleoncloud.org/experiment/share/. Last accessed 2025-02-20.
- 6 CloudLab. https://www.cloudlab.us/. Last accessed 2025-02-20.
- 7 SLICES-RI. https://www.slices-ri.eu/. Last accessed 2025-02-20.
- 8 Fed4Fire. https://www.fed4fire.eu. Last accessed 2025-02-20.
- 9 Sebastian Gallenmüller, Dominik Scholz, Henning Stubbe, Georg Carle, "The pos Framework: A Methodology and Toolchain for Reproducible Network Experiments," ACM CoNEXT'21, Munich, Germany (Virtual Event), Dec. 2021, https://dl.acm.org/doi/10. 1145/3485983.3494841.
- 10 SOSP 2023. https://sysartifacts.github.io/sosp2023/guide. Last accessed 2025-02-20.

4 Organisation of the Seminar

The seminar had an opening session with short (two-minute) presentations of all participants, in which participants not only presented themselves and their background in the area of the seminar, but also mention topics and questions they would like to discuss during the seminar. Plenary talks were followed by discussions on suitable topics for breakout working groups. For preparing the world cafe, in which in several rounds the participants give answers to a set of questions, suitable themes and questions were collected. Summaries from the breakout sessions and the world cafe rounds were presented and discussed in plenary sessions, followed by discussions on the key findings of the seminar.

5 Overview of Talks

5.1 SLICES, Research Infrastructure Defined as a Scientific Instrument

Serge Fdida (Sorbonne University – Paris, FR)

License ☺ Creative Commons BY 4.0 International license © Serge Fdida

A science is defined by a set of encyclopedic knowledge related to facts or phenomena following rules or evidenced by experimentally-driven observations. Computer Science and in particular computer networks is a relatively new scientific domain maturing over years and adopting the best practices inherited from more fundamental disciplines. The design of past, present and future networking components and architectures have been assisted, among other methods, by experimentally-driven research and in particular by the deployment of test platforms, usually named as testbeds. However, often experimentally-driven networking research used scattered methodologies, based on ad-hoc, small-sized testbeds, producing hardly repeatable results. We believe that computer networks needs to adopt a more structured methodology, supported by appropriate instruments, to produce credible experimental results supporting radical and incremental innovations. This presentation reports lessons learned from the design and operation of test platforms for the scientific community dealing with digital infrastructures. The SLICES initiative is introduced as the outcome multi-year process of conceptual evolution for a networking test platform transformed into a scientific instrument. Challenges, requirements and opportunities are addressed that our community is facing to manage the full research life cycle necessary to support a scientific methodology. Further details are provided in [1].

References

Serge Fdida, Nikos Makris, Thanasis Korakis, Raffaele Bruno, Andrea Passarella, Panayiotis Andreou, Bartosz Belter, Cedric Crettaz, Walid Dabbous, Yuri Demchenko, and Raymond Knopp, "SLICES, a scientific instrument for the networking community," *Computer Communications*, vol. 193, pp. 189–203, 2022. https://doi.org/10.1016/j.comcom.2022.07.019

5.2 Chameleon: New and Noteworthy

Kate Keahey (Argonne National Laboratory, US)

License ☺ Creative Commons BY 4.0 International license © Kate Keahey

An overview on the Chameleon testbed is given, with lessons learned and future enhancements. Chameleon provides bare-metal access to allow a high degree of configurability/isolation. It features adaptability, from large cloud testbeds to IoT. In addition to the two main sites at University of Chicago (UC) and at Texas Advanced Computing Center (TACC), there is a new core site at the National Center for Atmospheric Research (NCAR). Funding was recently extended by 4 years. Layer 3 connection between the sites at UC and TACC is realized through FABRIC. The Chameleon infrastructure is built on top of OpenStack. A catalog of experiments (Trovi) is available.

5.3 Orchestration for Reproducibility and Metadata Management

Sebastian Gallenmüller (TU München – Garching, DE)

Reproducibility is still an unsolved issue in the domain of computer science. To foster reproducible research, the ACM created badges that can be awarded to papers that provide high-quality artifacts allowing the recreation of experimental results. However, the badging system increases the effort for experiment providers and artifact evaluators. Despite the badges, reproduced experimental results may differ if seemingly minor details of an experiment are insufficiently documented.

To address these problems, we created the plain orchestrating service (pos), a framework that (1) orchestrates testbeds and (2) provides a reproducible experiment workflow. We provide a template for experiments that automates the entire experiment workflow from setup to measurement and evaluation. Live images ensure that all experiments start from a well-defined state. If experimenters adhere to the proposed workflow, the creation of reproducible experiments is ensured without additional effort for the experimenter or artifact reviewer. We call this property reproducibility-by-design.

Experiments that want to use the reproducibility-by-design feature depend on the availability of the pos controller. Previously, only our own testbeds at TUM run the pos framework natively. To increase the number of testbeds that can execute pos workflows, we ported the pos experiment controller to CloudLab and Chameleon. To run a pos experiment on CloudLab and Chameleon, we use testbed resources to deploy the pos framework. If a pos experiment requires 4 experiment nodes, we allocate 5 experiment nodes in CloudLab or Chameleon. The fifth node is used to temporarily host the pos controller that executes the experiment on the remaining nodes. After the experiment is over, the pos controller is removed, and the testbed can be used based on its original controller.

To increase the usability of the experimental results and foster the reuse of experimental data, we want to improve the way we describe and provide experimental results. RO-Crate is a standard that uses the JSON-LD format to define the experimental data of a research object and provide additional metadata, such as energy measurement data, the used OS images or software artifacts and their respective versions, or the topology of investigated

networks. These RO-Crates can be published on platforms like Zenodo to make them easily accessible and provide long-term availability. We plan to use RO-Crates as the default format for experimental results created by pos to offer a standardized way to bundle and archive experimental results.

We intend to provide the mentioned features, the reproducibility-by-design, the portability of experiment workflows between testbeds, and RO-Crate, a standardized way to bundle and provide research data for future SLICES testbeds.

5.4 Research Infrastructure Opportunities at NSF

Deep Medhi (NSF - Alexandria, US)

License $\textcircled{\mbox{\footnotesize \mbox{\footnotesize e}}}$ Creative Commons BY 4.0 International license $\textcircled{\mbox{$\mathbb O$}}$ Deep Medhi

This talk presents the viewpoint from the US research agency NSF on research infrastructures for networked systems research. A major reason why NSF makes a significant amount of funding available for research infrastructures is because these infrastructures enable new research. Concerning size of research infrastructures, the NSF nomenclature distinguishes:

- Major Research Instrumentation (MRI) (below 4M USD)
- $\blacksquare \text{ Mid-scale 1 (4-20M USD)}$
- Mid-scale 2 (20–100M USD)
- Large-scale (above 100M USD)

Current NSF-funded research infrastructures for networked systems research comprise FABRIC for connectivity (e.g., CERN, Hawaii, Chile) and compute resources; SAGE testbeds, a software-defined sensor infrastructure; EduceLab for Heritage Science; GMI3S for Internet security; SPHERE; Chameleon; CloudLab; NDIF (National Deep Inference Fabric) which addresses Software: AI inference libraries, Capacity: DeltaAI GPUs, and Education: interdiciplinary training on AI; PAWR: Wireless testbeds POWDER, Cosmos, Ara, Airpaw; Internet Measurement Research (IMR) Program for Methodologies, tools and infrastructure.

5.5 MERIF Mid-scale Experimental Reserach Infrastructure Forum

Paul Michael Ruth (RENCI – Chapel Hill, US)

License ☺ Creative Commons BY 4.0 International license ◎ Paul Michael Ruth

MERIF is presented, which was created to bring together the community to discuss and document compelling new opportunities for future NSF Midscale Infrastructures. Thereby, NSF supports midscale infrastructures that may help advance the state of the art, and build a stronger, more widely inclusive community. So far, the following MERIF workshops were held: MERIF Education Workshop (GWU, 2019), MERIF Future Experimental Research Infrastructures Workshop (FIU, 2020), MERIF Workshop (UW-Madison, 2022), MERIF Workshop (Boston University, 2023), and MERIF Workshop (UMKC, 2024). Participants of the MERIF workshops are associated with the following platforms: Cloud: Chameleon, Cloudlab; Networking: FABRIC; Wireless: COSMOS, POWDER, ARA, AERPAW; Security: SPHERE; Edge/AI: SAGE; AI: NDIF; Heritage Science: Educelab.

5.6 6G Visions and Possibilities for EU-US Collaboration on Network and Internet Technologies

Jorge Gasos (European Commission – Brussels, BE)

License O Creative Commons BY 4.0 International license O Jorge Gasos

An overview on research infrastructures and test platforms from the EU perspective is given in the context of SNS, the Smart Networks and Services, a Public-Private Partnership programme. The initiative is structured in three main streams: 6G technology research, experimental platforms, and large-scale trials with verticals.

The recently signed Administrative Arrangement and Partnership Plan between the European Commission – DG Connect and the NSF provides a framework for future EU – US research collaboration on network and internet technologies. In this context, future collaborations can be established for the experimentation, validation and benchmarking of 6G technology developments in research infrastructures and test platforms form the EU and US. In the area of internet technologies, collaborative research could address internet architectures, network security, trust and privacy, as well as electronic identities and decentralized technologies. While there is collaboration potential, details still have to be worked out.

5.7 Data Management Tools, Workflows, and Recommendations from NFDIxCS

Michael Goedicke (Universität Duisburg – Essen, DE)

License
 $\textcircled{\mbox{\scriptsize C}}$ Creative Commons BY 4.0 International license
 $\textcircled{\mbox{\scriptsize C}}$ Michael Goedicke

An overview of the National Research Data Infrastructure initiative (NFDI [1]) is given, with particular focus on the consortium of the National Research Data Infrastructure for and with Computer Science NFDIxCS [2]. The NFDI is a federally funded program organized by the Deutsche Forschungsgemeinschaft (DFG), with a budget of approximately 10–15M EUR for five years per consortium.

The NFDI operates through a bottom-up, discipline-oriented approach, and NFDIxCS addresses the specific sub-disciplines within computer science. NFDIxCS has 17 partners and additional participants, fostering a collaborative environment. International outreach is mandatory, ensuring global engagement and alignment with initiatives like the European Open Science Cloud (EOSC [3]).

The talk highlights several challenges in research data management, including managing diverse research data types, metadata, and software contexts within various research activities. To address these challenges, two central concepts are introduced: the Research Data Management Container (RDMC) and community participation supported by a central portal and services.

The RDMC is designed to create a strong link between actual research data, context information, metadata, execution environments, including the Software Bill of Materials (SBOMs). The RDMC provides a timecapsule which provides the facilities to bring the research data back to life through the aforementioned context information, software and a reusable execution environment. It references FAIR Data Objects [4] and RO-Crate [5] standards, emphasizing sustainability and interoperability.

Community participation is crucial for defining metadata standards, quality criteria for RDMC curation, and representation in international associations such as IFIP, Informatics Europe, and ACM/IEEE CS as well. This collaborative approach ensures that the subdisciplines of computer science are actively involved in shaping the NFDIxCS initiative.

In terms of RDMC and portal architecture / operation, the focus is on creating sustainable architectures, prototypes, reference implementations, and purpose-built versions. The emphasis is on reusable execution environments that can adapt to technological changes over time – in particular addressing also long-term archival.

Privacy and security are critical considerations. The talk outlines various levels of guarantees for data accessibility and privacy, ranging from open access with minimal guarantees to completely detached and standalone systems with no internet connection for maximum security and privacy. This includes the use of pseudonymized, encrypted, and role-based access control mechanisms.

The summary concludes by underscoring the initiative's commitment to fostering a collaborative environment for robust research data management practices in Computer Science and its sub-disciplines.

References

- 1 Nationale Research Data Infrastructure (NFDI). https://www.dfg.de/en/ research-funding/funding-initiative/nfdi. Last accessed 2025-02-20.
- 2 NFDIxCS. https://nfdixcs.org. Last accessed 2025-02-20.
- 3 European Open Science Cloud (EOSC). https://eosc.eu/. Last accessed 2025-02-20.
- 4 FAIR Data Objects. https://fairdo.org/. Last accessed 2025-02-20.
- 5 RO-Crate. https://www.researchobject.org/ro-crate/. Last accessed 2025-02-20.

5.8 FABRIC

Paul Michael Ruth (RENCI – Chapel Hill, US)

License $\textcircled{\mbox{\footnotesize \mbox{\footnotesize e}}}$ Creative Commons BY 4.0 International license $\textcircled{\mbox{$\mathbb O$}}$ Paul Michael Ruth

The Infrastructure Project FABRIC has a large number of Layer 1 fiber optical lines, which allow to create Layer 2 connections for users. FABRIC has an important number of international connections, including to Europe: Bristol, CERN, Brussels, and also to Japan: Tokyo. FABRIC has different connections to different testbeds, including Chameleon, CloudLab, Powdr, and also to cloud service providers, including AWS and Google Cloud. The architecture of FABRIC tries to accommodate different experimental needs, e.g., research in core networks, at the edge of the network, or in the cloud. FABRIC also offers access to real-world users and data, e.g., at experimental facilities at CERN, telescopes in Chile, or to P4 switches at different locations in the FABRIC network. This allows the creation of experiments for users using the FABRIC network as a connectivity provider or for interdisciplinary research between computer scientist and researchers from the application domain.

5.9 Post-5G Experiments

Damien Saucez (Inria – Sophia Antipolis, FR)

Traditional infrastructures primarily focus on providing access to computational and network resources, allowing researchers to deploy their own tools and conduct experiments. In contrast, SLICES-RI redefines this paradigm by placing the experiment at the core of its design. To achieve this objective, SLICES-RI is structured around the concept of blueprints, which serve as standardized frameworks facilitating collaboration and ensuring methodological consistency in experimental research.

Blueprints guarantee full understanding between engineers and domain researchers. They establish a unified terminology that does not necessitate deep technical expertise, thereby fostering an accessible yet rigorous approach to research design. By following a structured methodology, blueprints enable the selection of appropriate resources, iterative refinement, systematic validation, and the establishment of reproducible baselines for experiments.

Each blueprint is intrinsically linked to a specific research domain and is designed in direct collaboration with researchers in that field. This differs significantly from conventional infrastructures, which are often conceived and implemented outside of the research community they are supposed to serve. While this approach may seem to be not scalable, SLICES-RI ensures component reusability across different blueprints. Indeed, a reproducible experimental workflow, inspired by [9], has been embedded within SLICES-RI to standardize the research lifecycle.

The experimental process within SLICES-RI follows a structured sequence, beginning with a non-ambiguous definition of the experiment. Once defined, the necessary configurations and software components are automatically generated, removing manual intervention that may lead to reproducibility hazards. Subsequently, computational and network resources required for the experiment are provisioned within the infrastructure. With resources allocated, the experiment is executed under controlled conditions, leveraging both virtualized and physical hardware when needed. During execution, experimental results, metadata, system logs, event traces, input/output records, and additional telemetry, such as power consumption, are automatically collected and preserved. Finally, all gathered data and metadata are immutably packaged and published, ensuring that experimental results remain reproducible and verifiable.

We present the SLICES-RI Post-5G blueprint, which implements all these concepts in SLICES-RI. This blueprint incorporates both virtualized environments and real hardware, along with the software ecosystem to facilitate research across various dimensions of post-5G technologies.

Through engagement with the Post-5G research community, we identified what a good Post-5G blueprint had to provide. Researchers may utilize SLICES-RI as a comprehensive 5G environment for deploying verticals, without necessarily focusing on 5G itself, treating it as a research commodity. The blueprints also supports Software-Defined Networking (SDN) by allowing dynamic modification of network behavior, including the integration of custom xAPPs, which are essential for adaptive networking experiments. Furthermore, the blueprint allows network and radio modifications at the lowest layers to allows research on low level protocol layers. Additionally, the blueprint allows experiments with novel radio signal processing techniques, spectrum allocation strategies, and emerging antenna technologies such as THz frequencies to be conducted. Finally, Integration with High-Performance Computing

(HPC) is offered to address computationally intensive operations to be performed jointly with post-5G infrastructure, for example to run complex protocol fuzzing and AI-driven network optimization.

The early years of SLICES-RI development and pre-operational deployment have yielded significant insights into the design and implementation of experiment-centric infrastructures. Following a blueprint-based approach, where each blueprint is lead by the research community of the field, is a clear advantage of SLICES-RI, as it puts back the real objective of research infrastructures at the center: make researchers conduct their experimental research.

References

1 SLICES-RI community, *Slices-RI documentation*. https://doc.slices-ri.eu/. Last accessed 2025-02-20.

5.10 Joiner and NDFF

Andrew W. Moore (University of Cambridge, GB)

License $\textcircled{\mbox{\footnotesize \ensuremath{\varpi}}}$ Creative Commons BY 4.0 International license $\textcircled{\mbox{$\mathbb O$}}$ Andrew W. Moore

This talk gave an overview of two testbeds supporting networked-systems research in the United Kingdom. The two communications testbeds described are JOINER – a national infrastructure based upon UK-wide Layer-2 network, and NDFF – a UK dark fiber facility. Each testbed operates a network intended for research users to provide an evaluation environment that allows evaluations not otherwise possible in the academic domain.

Having been developing for over a decade, the National Dark Fibre Facility has over 1,300 km of single-mode optical fibre network, with control and monitoring systems. Access to the dedicated fibre network is provided at the physical layer through access points at four universities (Universities of Cambridge, Bristol, Southampton, and UCL in London). This permits experiments and research requiring *hands-on* access to the underlying physical media; such work has included evaluating new (photonic devices) receiver and transmission elements and lightwave amplification devices. Additionally, research into quantum communications pathways, such as quantum key distribution, requires direct access to dark fiber. The NDFF infrastructure enables this and has included the demonstration of world-record-length secure pathways. The NDFF Dark fibre infrastructure is a national facility underwritten by the UK Engineering and Physical Sciences Research Council (EPSRC) it has been operating in its current form since 2019 and has supported approximately 18 projects including next generation Internet, Sensing and Metrology, Terahertz communication, quantum communications, wireless communications and networking, and supporting the optical device, communications, and network communities. Alongside this, NDFF provides pathways that support layer-2 services and, in this way, is intended to coexist with the JOINER infrastructure collaboratively.

JOINER (Joint Open Infrastructure for Networks Research) is a project that uses new and existing communications paths to interconnect a standard rack of network infrastructure located at multiple sites across the United Kingdom to create a national layer-2 testbed. JOINER forms part of the UK Federated Telecoms Hubs, supporting the UK's strategy for future telecoms by providing joined-up infrastructure. Currently, eleven universities and national laboratories participate in hosting JOINER facilities. Once commissioned, JOINER is open to all, serving academia, entrepreneurs, government and enterprise.

The infrastructure consists of several general purpose compute, storage, and network resources and the provision of dedicated (re)programmable end-host hardware and network switching systems. At the time of this seminar, JOINER was scheduled for commissioning at the end of 2024. A long roll-out and multiple deployment sites have permitted some early JOINER experience, including demonstrating interoperability with the NSF-supported FABRIC project. JOINER has demonstrated connectivity with the FABRIC system through the node at Bristol as part of a demonstration for OFC 2024.

Questions for this talk highlighted the need for variety among testbeds: for example, variety in the *layer* at which experiments might be conducted and research enabled. An example is the need for dark-fibre access to perform experiments in the direct encoding of quantum photon pairs; some research experiments are impossible to undertake without appropriate access.

References

- 1 Joint Open Infrastructure for Networks Research, JOINER. https://joiner.org.uk/. Last accessed 2025-02-20.
- 2 National Dark Fibre Facility, NDFF. https://www.ndff.ac.uk/. Last accessed 2025-02-20.

5.11 SPHERE Research Infrastructure

Jelena Mirkovic (USC – Marina del Rey, US)

License $\textcircled{\mbox{\scriptsize \mbox{\scriptsize \mbox{\mbox{\mbox{\mbox{\scriptsize \mbox{\mbox{\mbox{\mbox{\scriptsize \mbox{\mbox{\scriptsize \mbox{\mbox{\mbox{\mbox}\mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\scriptsize \mbox{\mbox}\mbox{\mbox{\mbox{\mbox}\mbox{\mbox{\mbox{\mbox}\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox{\mbox\mbox{\mbox{\mbo}\mbox}\mbox{\mb}\mbox{\mbox{\mb}$

This talk gave an overview of the NSF-funded mid-scale research infrastructure project to build Security and Privacy Heterogeneous Environment for Reproducible Experimentation (SPHERE) [1]. The SPHERE research infrastructure will offer a novel mix of experimentation capabilities, uniquely tailored to the needs of cybersecurity and privacy researchers and educators. SPHERE's novel offering of diverse, rich hardware infrastructure, configurable network substrate and safe network policies will support novel cybersecurity and privacy research in emerging areas, such as IoT, cyber-physical systems, programmable networks, edge computing, Internet measurement, and human-centric cybersecurity. SPHERE's novel user portals will democratize access to cybersecurity and privacy research, and will facilitate practical education of broad student populations. SPHERE's novel support for representative experimentation and reproducibility, tight collaboration with researchers and close alliances with artifact evaluation committees, will enable vertical progress in the science of cybersecurity and privacy. The SPHERE research infrastructure aims to transform cybersecurity and privacy research, from piecemeal and opportunistic to highly integrated, by unifying the community's experimentation efforts on a common, rich, highly usable infrastructure. Jelena Mirkovic presented the SPHERE project's vision and elaborated on its status after the first year of the project.

References

1 SPHERE. https://sphere-project.net. Last accessed 2025-02-20.

5.12 European ESFRI Roadmap and Research Infrastructure Landscape (Digit Group)

Hakima Chaouchi (IMT - Palaiseau, FR)

The talk gave a brief introduction into the European funding scheme for large-scale research infrastructures, the European Strategy Forum on Research Infrastructures (ESFRI). Projects that pass the competitive ESFRI review process, will be included in the European ESFRI roadmap, which makes them eligible for specific funding calls.

The talk focused on research infrastructures from the European ESFRI DIGIT Group, explaining the landscape of the Data, Computing, and Digital Research Infrastructures domain. There are currently three ESFRI projects that are part of the DIGIT Group: SLICES, SoBigData, and EBRAINS. SLICES is the only ESFRI project that focuses on computer science research created by computer scientists for computer scientists. SoBigData addresses research in social sciences and EBRAINS neurological research. Further, a brief introduction of the French Research Infrastructure roadmap and landscape (RI Digital sciences) was presented.

6 Breakout Session

The seminar collected topics for the breakout session and discussed them. The identified topics agreed on for the working groups were:

- 1. Reproducibility
- 2. Workflows and Software in Research Infrastructures
- 3. Artificial Intelligence and Machine Learning Digital Twins
- 4. Testbeds and Science
- 5. Improving Testbed User Experience

6.1 Working Group: Reproducibility

Jelena Mirkovic (USC – Marina del Rey, US) Kate Keahey (Argonne National Laboratory, US)

The members of the working group were Yuri Demchenko, Sebastian Gallenmüller, Kate Keahey, Jelena Mirkovic, and Serge Fdida.

This group discussed goals of reproducibility initiatives, current pitfalls/gaps and how we could address them, and role of research infrastructures in supporting reproducibility.

Goals. Goals of artifact evaluation initiatives are seemingly to check the accuracy of authors' results and validate their claims. Another aspect of validation can be to test if the claims are robust enough that they hold on a different hardware or in a different environment. A better goal would be to enable researchers to "stand on the shoulders of the giants" i.e., to enable building on the prior work. This is called "practical reproducibility."

Role of Research Infrastructures (Testbeds). Testbeds facilitate practical reproducibility, because they enable testing of artifacts either in the same setting (e.g., if setup process is dependent on a given platform or OS) or in various environments (e.g., to test how environment changes impact results).

Challenges. One commonly mentioned challenge is incentives. Authors need incentives to share artifacts, evaluators need incentives to participate in artifact evaluation, and researchers need incentives to reuse existing artifacts instead of developing their own.

Author incentives. Funding agencies can either offer additional small funding for artifact publication or mandate open science. Some data may contain sensitive information and may require a lengthy approval process to sanitize and publish – not all artifacts are equally valuable or equally easy to publish. For example, a dataset may contain data collected in a testbed experiment and thus easily releasable, or data observed on the Internet, which usually has to be anonymized for release. Anonymization and required approvals for release (e.g., IRB approvals) add cost to observational dataset release.

We do not want to create a climate, where unstructured data is published, just to meet the mandate. Instead, publication should be rewarded, but not required. Some quality control is necessary (e.g., via artifact validation) to ensure that published artifacts are useful and usable. Tenure and promotion committees could also help incentivize artifact authors by counting artifact publications towards promotion.

Evaluator incentives. Artifact evaluators are currently not adequately rewarded. Participation in AE committees could help students get into graduate school, but publications still carry significantly higher weight. If conferences offered free registration or travel grants for evaluator service, this would boost evaluator recruitment.

Artifact Standards. To improve reproducibility we should standardize artifact metadata. Some of that metadata can be extracted automatically from artifact repositories. It would also be very useful if we could develop automated checkers to detect if some data or code dependency is missing from the artifact. Testbed operators can provide support for data export about resources and their setup when an artifact is packaged on the testbed.

Another need is for the research community to come up with artifact quality standards. This would enable researchers to understand how valuable an artifact may be for their research, and how easy it would be to reuse it. Datasets face some unique challenges around measuring quality and understanding usefulness, since usefulness of datasets depends on a researcher's goal. Citations are not a good measure of dataset quality. In some cases, the dataset may be full of errors, or outdated, but it may be the only dataset of the given type, and thus widely used by the community.

6.2 Working Group: Workflows and Software in Research Infrastructures

Damien Saucez (Inria - Sophia Antipolis, FR)

License $\textcircled{\mbox{\scriptsize \ensuremath{\varpi}}}$ Creative Commons BY 4.0 International license $\textcircled{\mbox{$\mathbb{O}$}}$ Damien Saucez

The members of the working group were Tom Barbette, Michael Goedicke, Raymond Knopp, Deep Medhi, and Damien Saucez.



Figure 1 Structured experiment workflow.

This breakout group focused on defining workflows and software needs for research infrastructures to ensure reproducible experimental research. Two key assumptions were established: (i) correctness of experiments remains the responsibility of the researcher, and (ii) infrastructure policies are expected to be followed, with logging and monitoring serving as diagnostic tools rather than enforcement mechanisms.

Requirements for Experiment Workflows. A robust experiment workflow must support essential research activities, including peer review, publication, artifact validation, and simplifying the review process, particularly for artifacts and research data.

Collaboration in experiment workflows must facilitate shared contributions among multiple researchers, enabling concurrent work on different components. Experiments should be composable, allowing sub-experiments to be interconnected. For example, a teacher could deploy a 5G core while students attach their own radio networks to it.

Data access policies must distinguish between internal sharing among experiment participants and external dissemination. This raises questions regarding authorship and ownership in multi-party workflows. Additionally, experimental research extends beyond final publications—many experiments serve exploratory purposes, requiring iterative trial-and-error processes. Workflow design should accommodate features such as pausing experiments and creating checkpoints, akin to CI/CD pipelines and interactive execution modes. Overly restrictive workflows may drive researchers toward ad hoc solutions of their own, reducing reproducibility.

A structured workflow, inspired by [9], has been proposed (cf. Figure 1).

In shared infrastructures, complete isolation is unattainable. Testbeds must transparently define their isolation model such that researchers can determine if it is suitable for their experiments. Diversity in testbeds and software stacks is inevitable (and wanted), given the varying requirements across research domains. Ensuring portability requires balancing comprehensive custom solutions with the integration of modular, off-the-shelf software components.

Reproducibility in testbeds may lead to the ossification of the testbed: modifications to testbed software may alter experimental results. To mitigate this, clear documentation on potential behavioral changes and migration strategies is essential. Using up-to-date software will attract users, particularly students transitioning to industry, who benefit from familiarity with widely used tools (e.g., Ansible, Terraform, GitLab pipelines as of 2024) rather than proprietary testbed-specific solutions. Ensuring long-term reproducibility remains an issue.

Regression testing presents a viable solution: whenever the testbed undergoes changes, previous experiments should be re-executed to assess how the changes impact the results. Testbeds may also introduce certification mechanisms, such as badges, to validate that results were obtained using a specific infrastructure configuration.

Finally, result publication should transparently document the experimental environment. A systematic section in research papers acknowledging infrastructure providers could enhance recognition and reproducibility of experimental research.

6.3 Working Group: Artificial Intelligence and Machine Learning Digital Twins

Georg Carle (TU München – Garching, DE) and Walter Willinger (NIKSUN – Princeton, US)

 $\begin{array}{c} \mbox{License} \ensuremath{\,\textcircled{\textcircled{}}}\xspace{\ensuremath{{}}\xspace{\ensuremath{\}}\xspace{\ensuremath{}}\xspace{\ensuremath{}}\xspace{\e$

The members of the working group were Georg Carle, Tobias Hoßfeld, Wolfgang Kellerer, Andrew W. Moore, Walter Willinger, and Martina Zitterbart.

The working group was set up to address the following questions: What needs to be considered when creating Digital Twins by means of AI-based Data and Workflow management? Does the development of next-generation Foundation Models (FM) for networking lead to a new kind of "collaborative networked systems research"?

The working group started with analyzing the creation of a Digital Twin within a testbed. A guiding example for which the creation of a digital twin would be desirable is *autoscaling* (scale-out/scale-in) of a Kubernetes Network Function Virtualization (NFV), e.g., as part of a 5G/6G use case. The digital twin would allow to analyze the behavior of the autoscaling function for different input and different configuration parameters. In general, to create a digital twin for a given task, there exist different approaches. One important traditional approach is to create a (hand-crafted) simulation; a more recent approach leverages Machine Learning (ML) and Artificial Intelligence (AI). When applying AI/ML, different AI/ML methods can be used, in particular:

- 1. black-box learning (e.g., supervised learning),
- 2. grey-box learning (e.g., explainable AI), and
- 3. Foundation Models (FM) that use self-supervised learning and unlabeled data for pretraining (phase 1) and then perform application-specific finetuning (phase 2).

In order to assess the different approaches of generating a digital twin, it is important being able to assess the prediction quality (concerning what-if questions) of a digital twin. Scientific questions in this context include:

- What is the "right" training data for a given problem?
- What are possible root causes of inaccurate or wrong inferences? (E.g., are root causes due to a specific machine learning method used, or due to bad training data?
- Decision tree (DT) models are appealing because they are inherently interpretable, but does their simplicity limit their applicability?
- Do DT models miss out on important hidden context?
- How to quantify the confidence in the output of a trained DT model?
- How to test for model quality beyond accuracy?
- Is the generated model generalizable? (An example for a non-generalizable model would be that the prediction is only accurate for a specific testbed, which was used to generate the data.)
- What is a good FM?

Domains for which such digital twins could be generated for answering scientific questions include performance and security. With regard to model quality, mediocre models will be able to answer specific questions with sufficient accuracy, while we can expect that from a good model, a domain expert can learn from its outcomes. While large FMs may have the potential to be a basis for good models, realizing them using currently employed workflows or pipelines poses significant challenges, including high costs (e.g., specialized hardware), high energy consumption (e.g., computing power requirements in the terraFLOPS range), and long model training times (e.g., days or weeks). Other open problems associated with FMs include whether different FMs need to be generated for different domains (e.g., for wired networks, for wireless networks, for cybersecurity, etc.), and how to create good FMs (e.g., by creating hybrid models that include domain-specific hand-crafted models). With respect to good datasets, testbeds are considered valuable (i.e., necessary but not sufficient) to generate realistic data sets.

6.4 Working Group: Testbeds and Science

Hakima Chaouchi (IMT – Palaiseau, FR) and Walid Dabbous (INRIA – Sophia Antipolis, FR)

The members of the working group were Hakima Chaouchi, Walid Dabbous, Bamba Gueye, Tobias Hoßfeld, Björn Scheuermann, Henning Schulzrinne, and Jörg Widmer.

The working group started looking at testbed coverage and futures, discussing whether the current set of testbeds provides all necessary scientific instrument functionalities. The discussion covered gaps in current testbeds, emerging use cases, and the role of testbeds in advancing research. Participants explored testbed categorization, research question formulation, and the potential alignment with methodologies used in other scientific disciplines.

A scientific instrument is a research infrastructure designed to enable systematic experimentation and validation of scientific theories. In digital sciences, this means testbeds should go beyond simply demonstrating technology – they must support rigorous experimentation, reproducibility, and fundamental scientific inquiry, similar to physics research facilities.

There are two approaches to research testbed design in computer science:

- Testbed-Driven Research: Researchers use available testbeds and define research questions accordingly.
- Research-Driven Testbed Design: Research questions are defined first, and testbeds are designed or selected based on scientific needs.

The working group emphasized the importance of the second approach, similar to physics research, where experimental facilities are built to answer fundamental scientific questions rather than just demonstrate technology feasibility.

The group identified three main categories of existing testbeds:

- 1. General-Purpose, Flexible Testbeds Adaptable platforms for a wide range of research topics.
- 2. Domain-Specific Testbeds Built for targeted experimentation in realistic but controlled environments, including simulation & emulation.
- 3. Unique Capability Testbeds High-cost infrastructures supporting specialized experiments (e.g., quantum networks, satellite communications).

These testbeds have several gaps in their functionalities:

- Scalability: Large-scale experiments remain challenging due to cost and infrastructure limitations.
- Data Accessibility: Many research questions, especially in AI-driven networking, require real-world data that is not publicly available.
- Technology Evolution: Rapid advancements in 5G/6G and AI make it difficult for testbeds to remain relevant over time.
- Reproducibility and Observability: Unlike in physics, digital testbeds often struggle with experiment reproducibility and privacy-related constraints.
- Reconfigurability vs. Stability: Some research requires testbeds to be highly modular, while others demand a stable and reliable setup.

Moreover, several research challenges demand improved testbed capabilities:

- AI-driven network automation: requires testbeds with real-world telecom data and high reliability.
- Digital twin-based testbeds: essential for replicating complex, real-time network conditions.
- Open, decentralized networking research: enables investigation beyond industry-driven technologies.
- Large-scale software emulation: facilitates cost-effective scalability for networking experiments.

Based on the discussions the working group made the following recommendations:

- 1. Testbeds should be designed to serve as scientific instruments, ensuring they support: (1) systematic experimentation and theory validation, (2) reproducibility and rigorous scientific inquiry, and (3) scalability and adaptability to future research challenges.
- 2. Develop a research-first methodology for testbed design guiding researchers to (1) define research questions independently of available testbeds, (2) identify the necessary testbed characteristics (e.g., scalability, modularity, real-world data integration), then (3) determine whether existing testbeds can be adapted or if new infrastructures are needed.
- 3. Establish large-scale and open research testbeds; given industry constraints on testbed access, academia should explore independent initiatives such as:
 - "Build Your Own Network" (BYONET) A modular, open-source testbed for networking research, free from industry limitations.
 - Academic Cellular Network for AI Research A dedicated testbed for large-scale telecom data collection and AI-driven networking studies.
 - Scalable Software-Based Emulation Tools that allow large-scale experiments without requiring physical infrastructure.

The working group concluded that while existing testbeds provide essential functionalities, significant gaps remain in scalability, sustainability, and alignment with emerging research needs. Addressing these challenges requires better research question formulation, modular and reconfigurable testbeds, large-scale academic-led initiatives, and closer collaboration with industry. By adopting structured methodologies and leveraging lessons from physics and engineering disciplines, the research community can ensure testbeds continue to serve as powerful scientific instruments for future innovations.

6.5 Working Group: Improving Testbed User Experience

Terry Benzel (USC - Marina del Rey, US) and Paul Michael Ruth (RENCI - Chapel Hill, US)

License ⊕ Creative Commons BY 4.0 International license © Terry Benzel and Paul Michael Ruth

The working group was focused on a range of topics on Improving User Experience (with testbeds). The participants were Terry Benzel and Paul Ruth – organizers, Jim Kurose and Adam Wolisz.

The initial topics for exploration included testbed evaluation, testbeds in the research lifecycle, i.e., how to make them a place where new technology is developed as opposed to mostly testing and evaluation of semi-mature research technology; eliciting input from the testbed user community early in the development of testbeds, i.e., how do testbed users and developers come together on needs for testbed services; how to balance research versus educational use of research infrastructures, identifying examples of specific types of testbeds in hardware and communications (5G/6G); balancing testbed access models on-site versus remote and sustainable testbeds and testbeds for sustainable IT. Note not all topics were discussed, given time limitations.

The discussion began by suggesting that the community needs a "catalog" of testbeds and some means of identifying features, and by posing the question if there is a way to loosely standardize or create an *ontology of testbeds* and can this lead to approaches to common (but not mandated interfaces). Some concerns were raised about the term "ontology" as possibly being too prescriptive.

A more general discussion ensued around defining models that could be reusable on arbitrary testbeds, though that might be too ambitious. To realize this vision, there is a need for *basic building blocks that can be easily tied to sets of research problems*. These need to be structured through *layers of abstraction*. Abstractions are best created by user communities for specific use cases. One approach to developing packages is to crowd-source through user-packaged and *shared experiments*. Shared experiments should include scenarios and stories that show how to create an experiment using these examples. As always, there are challenges in bridging user communities from the novice to the expert users. Each may need different building blocks and scenarios. There is no one-size-fits-all. Some newer research infrastructure projects are developing different use "portals" for the different user communities. To scale, there is a need for large project use cases integrated with people, processes, and design experts. These are not trivial recommendations and typically go beyond the scope of most research infrastructure project support.

The group then turned to the question of how research infrastructure developers/operators can *attract user communities*. Understanding who the users are and matching research infrastructure to different communities is important. Some are interested in low-level technologies, while others are designing (or using) applications. This is related to the initial discussion about building blocks, layers of abstraction, and ontology.

At present, each testbed engages in "traditional" approaches to outreach, including presenting at conferences, recruiting users to present at conferences, and a certain level of evangelism. Ideas to shift these approaches to more coordinated, interactive, and high-visibility recruitment include competitions, grand challenges, and laboratories for education classes. Recruiting, maintaining, and growing user communities is vital to research infrastructure success. It is strongly recommended that testbed funding should include resources for creating and marketing users and educational exercises/experiences.

The final point of discussion was an exploration of roles for research infrastructure/testbeds outside of research labs and educational institutions. In these cases, the demand needs to come from the organizations, but this becomes a catch-22: if they don't know about research infrastructure, how can they understand how it fits in their workflow? In general, R&D organizations in major tech companies have or can develop research infrastructure. On the other hand, companies further down the tech innovation chain may not have cycles to develop and operate exploratory research infrastructure. We recommend adopting initiatives to extend publicly funded research infrastructure through private funding while maintaining both open and proprietary use. There are challenges in moving from testbeds as exploratory tools, to mature technology transfer and production systems. One approach to balancing these competing demands is to create a research infrastructure as a service model.

7 World Cafe

In this session, participants discussed six themes in the format of a world cafe. The themes were identified from brainstorming among the participants. Each theme was assigned a moderator and a specific room. Participants rotated between the different rooms, each dedicated to specific theme. The moderator led the discussion and summarized the results. The following themes were discussed:

- 1. 5G/6G and 3GPP
- 2. Testbed Evaluation
- 3. Long-running Experiments
- 4. Better Testbeds
- 5. Sustainability
- 6. International Collaboration

7.1 World Cafe Theme: 5G/6G and 3GPP

Jörg Widmer (IMDEA Networks Institute – Madrid, ES)

License $\textcircled{\textbf{G}}$ Creative Commons BY 4.0 International license $\textcircled{\textbf{O}}$ Jörg Widmer

The discussion first focused on the issue that it is very difficult to run experiments with 5G/6G industrial requirements on academic research infrastructures, and how academia and industry can collaborate on 5G/6G research. One of the main issues is that the capabilities of testbeds built by academic researchers lag behind the capabilities even of current production RAN systems, let alone provide a platform for research on future generations of mobile networks. Academia thus struggles to access high-quality 5G/6G hardware for experiments. Most industry research in that area takes place behind closed doors. Open platforms like SLICES exist, but given their characteristics they are of limited use for industry. It is a very difficult question how to design testbeds that industry will actually use. Some operators, like Deutsche Telekom, are open to working with universities, but vendors are more hesitant because of confidentiality and proprietary technology. ITU-T has proposed distributed testbeds, but they are complex and hard to implement. That's why many researchers rely on simulations, emulations, or small private 5G setups with OpenAirInterface (OAI) as a more practical solution.

While a potential idea for academia is to work with industry on industry-internal testbeds, this is not very realistic due to the sensitivity and confidentiality of such testbeds, and comes with many challenges like NDAs and limited access to new technology. One way forward is to form research partnerships and public-private collaborations to align research goals. A possible approach is to create shared testbeds that both academia and industry can use, with realistic network traffic to bridge the gap between experimental and real-world scenarios.

The second topic discussed was the role that universities can play in standardization, and specifically in 3GPP. 3GPP standardization is almost entirely controlled by industry and is highly driven by "political" and economic aspects rather than purely technical, making it hard for universities to contribute. Unlike Internet standards, which evolve through open collaboration, 3GPP operates largely behind closed doors. Institutions like Fraunhofer and EURECOM are involved, but most universities are not. Timing is a big issue—by the time academic research produces results, industry has usually moved on. Many academic testbeds are already outdated by the time they could influence the standard. To make an impact, researchers would need to look beyond current 5G setups and focus on what comes next, which is difficult. One possible area of contribution is the higher-layer Service Architecture (SA), especially in network automation, AI-driven management, and security. Open RAN (O-RAN) could also be an opportunity, since it promotes modular and flexible network design, making it easier for academic research to have an impact.

The discussion made it clear that it would be desirable for academia and industry to collaborate more closely, with better access to open testbeds and a stronger effort to keep research relevant to future 5G/6G developments. However, there are significant obstacles. Universities need to engage with industry early, work together on research plans, and align with standardization timelines. Public funding can help to some degree by supporting joint projects and open testbed initiatives. However, ultimately the biggest hurdle is the reluctance of industry to collaborate due to confidentiality concerns, the difficulty to match timelines and oftentimes unclear direct benefit of such collaborations. Academia needs to show industry why its research matters and find ways to integrate it into the development process.

7.2 World Cafe Theme: Testbed Evaluation

Georg Carle (TU München – Garching, DE)

License

Creative Commons BY 4.0 International license

C Georg Carle

The discussions of this theme had the goal to identify suitable criteria to evaluate and compare different testbeds. This includes the question of what are good metrics. Additionally, it was discussed how to evaluate plans for future testbeds.

A number of quantifiable criteria can be used to measure and compare the performance of its components and amount of resources, economic value, impact and quality of different testbeds. One also has to be aware that testbeds are highly diverse and serve different communities, which partly have largely different sizes, so quantifiable criteria must be used carefully.

While measuring the amount of resources and the associated economic value means using a limited number of well-understood metrics, a fairly large number of metrics can be used for measuring impact and quality.

One aspect of impact are scientific results. This includes the list of huge discoveries, which, however, is frequently empty. Relevant metrics to measure scientific results are the number of publications that cite a testbed (however, one must be aware that publication may be much later than testbed usage), the number of of citations of these publications (however, citations may take up with a long delay). From these citations, it is possible to calculate an H index for testbeds (while its usefulness is debatable).

Another aspect of impact relates to usage of a testbed. Metrics include the number of single users (active per month and cumulative), the number of organizations to which users belong, the number of experiments (active per month and cumulative), user retention (measuring the share of users that stick around), and "monthly active users" (MAU), which is a standard statistic for web platforms. The number of users can be related to the size of community addressed by a testbed, to derive the usage in percent within that community. Usage also relates to education activity, where suitable metrics are the number of educational events, and the number of students trained. Usage also relates to industry involvement, with number of industrial users; number of SMEs, and percentage of industrial usage as suitable metrics.

Another aspect of impact relates to artifacts available in the context of a testbed. Metrics include the number of software artifacts, including contributions to open-source (cf. GitHub metrics), and the number of experiments available. Further metrics can be derived to assess standardization relevance, based on supported standards (evaluation and prototyping), and the number standard contributions with artifacts in the testbed.

To assess the quality of a testbed, usability metrics can be used. An innovative usability metric can be defined by measuring TTFE: Time to first experiment, i.e., how much time does an average user need to realize a representative experiment.

As testbeds are highly diverse, different different categories of testbeds can be defined. It was concluded that different sets of evaluation criteria are needed for different categories.

7.3 World Cafe Theme: Long-running Experiments

Sebastian Gallenmüller (TU München – Garching, DE)

Long-running experiments were the main topic of this discussion group. Experiments in this category typically run for weeks or months. Typical platforms for such experiments were the now defunct PlanetLab [2], or current platforms such as Ripe Atlas [3] and FLOTO [1].

We identified three areas for long-running services on testbeds: (1) observational studies, (2) monitoring, and (3) research infrastructure as a service provider:

Observational studies monitor networks, such as the Internet. The main purpose of these experiments is to record data for research. Typical parameters to observe are the performance of services, including temporary or long-term changes in performance, the reliability or availability of services, their energy consumption, or the human interaction with these services.

Monitoring is a different use case for testbeds. The observed parameters are similar to observational studies; monitoring applications typically check the availability and potential performance degradations of services. However, the main purpose of the data collection is not research but maintenance. Testbeds can be used to monitor internal and external purposes,

i.e., to observe the services of the testbed or to monitor services hosted outside the testbed. Despite of the original focus on maintenance, the collected data may eventually be used for research.

Research infrastructure as a service provider uses the facilities of a testbed to host services. A simple service of a testbed-run service would be a webserver. Another notable example of such a service was CoDeeN [4], a content delivery network hosted on the PlanetLab testbed.

We further identified requirements typical for long-term experiments. The computing and bandwidth resources that are consumed by these experiments are typically limited. The energy consumption and the storage requirements for data recording can be significant. Long-term experiments can prevent the testbed from being switched off, increasing its power consumption. Specific APIs may be needed for specific measurement tasks, e.g., for energy measurements. At the same time, shared resources must be managed to ensure a fair resource distribution between the experiments running on a testbed. Based on these preconditions, we deducted recommendations for testbed providers to host this type of experiment.

A major difference concerns the *usage of resources*. Observational studies run over weeks and months, monitoring services or services on research infrastructures may even run for years. The runtime of these experiments may be even longer as researchers extend measurement times to increase the amount of collected data. Another typical usage pattern is periodically scheduled experiments that run their task according to a service schedule. To handle such experiments, the usage policy of a testbed should be defined clearly. A possibility to minimize interference with other experiments is an automated scheduler. Such a scheduler can shift execution times to less attractive time slots where the testbed usage is lower. Another recommendation is to make the experiments reproducible. The process of reproducing experiments should be fully automated to make it as easy as possible [5]. This automation can then be used to recreate the measurement process on the testbed without needing the assistance of the experimenters.

The nature of long-running experiments requires a different approach to *data validation*. Typically, data is validated at the end of an experiment, when data collection is complete. However, for long-running experiments, the collection process may never be complete. Therefore, we recommend shifting towards an ongoing validation process that constantly or periodically checks the validity of collected and processed data. This validation process must notify the experiment in case of an error during the runtime of an experiment. We further recommend informing the experimenter about the resource usage or the underutilization of resources. This serves two purposes: it incentivizes responsible use of resources and helps identify potential errors.

The observational *data* measured by long-term experiments depends on the current state of the observed system. This state of the system, e.g., the Internet, is constantly changing, making the observed data unique for a specific point in time and vantage point. The data can not or not easily be recreated, which gives an inherent value to the measured data. The analysis of data and publications based on this analysis create additional value for scientists that is recognized and incentivized in the scientific community. However, this process prevents the sharing of data before a publication is accepted. To mitigate this problem, we recommend that high-profile conferences and journals should accept publications on data sets to motivate people to share data as early as possible. We further recommend sharing intermediate data sets that can be extended over time.

References

- Alicia Esquivel Morel, Mark Powers, Kate Keahey, Zack Murry, Tomas Javier Sitzmann, Jianfeng Zhou, and Prasad Calyam. FLOTO: beyond bandwidth – A framework for adaptable, multi-sensor data collection in scientific research. In *High Performance Computing. ISC High Performance 2024 International Workshops – Hamburg, Germany, May 12-16,* 2024, Revised Selected Papers, volume 15058 of Lecture Notes in Computer Science, pages 427–438. Springer, 2024.
- 2 Larry L. Peterson, Andy C. Bavier, Marc E. Fiuczynski, and Steve Muir. Experiences building planetlab. In 7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8, Seattle, WA, USA, pages 351–366. USENIX Association, 2006.
- 3 RIPE Ncc Staff. Ripe atlas: A global internet measurement network. *Internet Protocol Journal*, 18(3):2–26, 2015.
- 4 Limin Wang, KyoungSoo Park, Ruoming Pang, Vivek S. Pai, and Larry L. Peterson, Reliability and Security in the CoDeeN Content Distribution Network. In Proceedings of the General Track: 2004 USENIX Annual Technical Conference, June 27 – July 2, 2004, Boston Marriott Copley Place, Boston, MA, USA, pages 171–184. Usenix, 2004.
- 5 Kate Keahey, Jason Anderson, Mark Powers and Adam Cooper. Three Pillars of Practical Reproducibility. In 19th IEEE International Conference on e-Science, e-Science 2023, Limassol, Cyprus, October 9-13, 2023, pages 1–6, IEEE, 2023.

7.4 World Cafe Theme: Better Testbeds

Björn Scheuermann (TU Darmstadt, DE)

This theme had the goal to identify how to get to better testbeds. The discussions focused on a number of key aspects, recognizing that "better" is a term with many facets.

First, "better" may refer to the usability of the testbed. The *time to first experiment* early on emerged as a key term in the discussion, and was quickly taken up by several later world cafe groups. It emphasizes the need for low entry barriers for new, first-time testbed users. This can, for instance, mean an authentication infrastructure that spans multiple testbeds (and/or other research infrastructures), continuous admin support and the availability of trained experiment support staff, and a community building concept that creates opportunities to learn, interact and get support among testbed users. More generally, future testbeds will need to overcome potential early users' commonly observed skepticism about infrastructures that they haven't worked with yet and where they didn't participate in the infrastructure design – in the discussion referred to as the "not invented here syndrome."

"Better" can also refer to the fit of the testbed for the specific experimental purpose. There is a wide range of entirely different research infrastructures, which have been referred to by the term testbed, from highly purpose-tailored, very specifically instrumented setups to very generic compute infrastructures. They vary largely with respect to which types of experiments they can support. Testbeds for networking research also vary with respect to the degree of isolation that they provide: shielded lab setups versus research infrastructures across the Internet, capturing real-world influences from shared use and real cross traffic. The latter may provide additional insights, but at the same time limits repeatability. There was agreement that, due to the largely varying demands of experiments, there cannot be a one-size-fits-all testbed infrastructure. The discussion did, however, also stress that there

often is a transparency deficit with respect to what a specific testbed can do versus what it cannot do, making it very hard to find the "right" testbed for a planned experiment among those that are available. A testbed database with clear information what can be done on and what can be expected from each specific testbed was mentioned as a potential step forward.

Building "better" testbeds in the light of these insights requires to be in touch with the potential user community, to understand what is actually needed (and cost efficient). It also demands to build on experiences from previous testbeds, including the lesson that too tight budget constraints can result in infrastructures that are hardly accessible due to usability deficits and flexibility deficits, and hence hardly used. Because components for cutting-edge research tend to be short lived, continuous evolution of testbed infrastructures and a lifecycle model need to be considered from the beginning, to improve the ratio between buildup time and actual operational time.

Finally, "better" testbeds can also mean testbeds that are more useful to the community. This obviously interrelates with the previously mentioned dimensions of "better." But it also includes aspects like automated documentation of testbed experiments to encourage and support independent repeatability of experiments, the searchability of results obtained through the testbed, or follow-up use of phased-out testbeds for instance for educational purposes.

7.5 World Cafe Theme: Sustainability

Tom Barbette (UCLouvain, BE)

License $\textcircled{\textbf{O}}$ Creative Commons BY 4.0 International license $\textcircled{\textbf{O}}$ Tom Barbette

The discussion focused mainly on environmental and societal sustainability, mostly leaving over the economic sustainability of testbeds. This was mostly due to time limitation, as we note they are highly interconnected as an environmentally friendly solution, which is not viable economically, is worthless. The actionable propositions below are, however, generally cheap. It is also to be noted that although some research in itself involves a large computational effort, the research outcomes might have highly beneficial rewards on sustainability. The impact of testbeds as defined on the Sustainable Development Goals is to put in balance with their benefits.

Efficiency of the shared infrastructure. As for cloud computing, using a shared infrastructure generally enables a better efficiency than scattered, under-utilized equipment. For instance, the PUE of centralized testbeds is generally better than the local closet individual teams generally use.

Life-cycle analysis for better environmental sustainability. It was noted that hardware manufacturing accounts for roughly half of total environmental impact of servers [2], which is often unaccounted for, with users generally focusing on energy.

The testbed interface could therefore help the user diminish its carbon impact in a number of ways, such as indicating a global CO_2e account for the experiment after the completion of the experiment. For instance, a message could be displayed to the user such as "Your experiment emitted X kg of CO_2e ." Accounting for different possible deployments, the interface could also propose multiple scenarios, multiple possible hardware, allowing the user to select the most sustainable deployment according to various tradeoffs.

This effort could be rewarded through a new paper KPI, allowing claims such as "my new algorithm has X% better accuracy and also trains using half the CO₂e on this standardized testbed." Further work is needed for selecting the right metrics such as the energy for the value of the experiment rather than a purely quantitative metric, whether the local carbon intensity should be used rather than global mix, which one should be selected to allow the aforementioned comparison when experiments were run at different time, how to get the consumption of individual component, ...

The discussion also tried to gather why machines always stay on in practice. Some participants explained turning on and off machines repeatedly increase the failure rate, long booting times that prevent high utilization of the testbed, and a tendency for overprovision from the users. Users themselves overprovision because they want to keep the state of the machines, or keep the reservation itself. It is noted that bare-metal providers reboot machines for each experiment anyway. Chameleon notably powers off machines between experiments. A possible mean of action could be to **enable a credit-based system**, even if infinite in practice as users could request easily more credits, the psychological effect might be enough to incentivize users not to hold unused (and therefore powered on) machines.

Finally, regulations such as done in ESFRI, can **mandate a policy on waste manage-ment**. More research could also be done on practical evaluation of whether hardware should be renewed or keeping less efficient hardware for longer. Experiment testbed possess data that might be leveraged to refine the leverages on this tradeoff. Further work could define an optimization problem to decide whether some machines should be turned on/off between experiments or not.

Efficiency of the experiments. Better testbeds, mean better usability and therefore better workflows. By **integrating a support for experimental design** such as discussed in Section 6.2, one can reduce the number of unneeded exploration of parameters and over-repetition of experiments. The testbeds could also incentivize to use emulation/simulation/virtualization when it is sufficient. For instance enoslib[1] enables to use the same script on a virtual infrastructure or a physical infrastructure, allowing most functional evaluation to be executed on the former and final evaluation on the later.

Societal sustainability. The testbeds also play an important role on education and access to expensive resources for the masses. It is important to develop a narrative towards places where access to testbeds, equipment and/or education is more difficult.

Future research. Beyond the aforementioned research directions, it is noted projects such as GreenDIGIT [3] already address similar concerns on 4 research infrastructures: EGI, SLICES, SoBigData, and EBRAINS and will investigate some of those proposals.

References

88

- Ronan-Alexandre Cherrueau, Marie Delavergne, Alexandre Van Kempen, Adrien Lebre, Dimitri Pertin, Javier Rojas Balderrama, Anthony Simonet, and Matthieu Simonin. Enoslib: A library for experiment-driven research in distributed computing. *IEEE Transactions on Parallel and Distributed Systems*, 33(6):1464–1477, 2021.
- 2 Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pages 854–867. IEEE, 2021.
- 3 Horizon Europe EU program. Greendigit. https://cordis.europa.eu/project/id/ 101131207.

7.6 World Cafe Theme: International Collaboration

Paul Michael Ruth (RENCI – Chapel Hill, US)

License O Creative Commons BY 4.0 International license O Paul Michael Ruth

This theme had the goal to identify what can be done to foster collaboration internationally.

It was discussed to restart the series of Global Experimentation for Future Internet (GEFI) workshops that brings together the international testbed community. This workshop could be used to share tooling and educational material between testbeds, and involve new communities, e.g., from DIGITAfrica.

Established platforms such as MERIF or SLICES could offer grants to foster graduate student exchanges, e.g., for one semester or a summer internship. However, such exchange programs will only work if funding is provided through the respective funding agencies.

A federation between testbeds was discussed. This federation should include the sharing on several levels, such as, a logical sharing of information, e.g., to create shared accounts across testbeds, or share data between testbeds.

Testbeds could collaborate to offer joint support of reproducibility efforts at major conferences. Different testbeds offering a diverse set of capabilities and resources will be needed to accommodate the resources required for artifact evaluation at major conferences.

To promote the testbeds, their capabilities, and benefit for the scientific community, we should use existing channels, such as the networking channel [1] or 6Gxcel [2] to make them more known among our community.

A global collaboration between testbeds could extend the capabilities of testbeds by adding complementary resources and experiences, add novel features not present in other countries, or regulatory advantages, e.g., due to different policies regarding privacy.

References

- 1 The Networking Channel. https://networkingchannel.eu/. Last accessed 2025-02-20.
- 2 6G-xcel. https://www.6g-xcel.eu//. Last accessed 2025-02-20.

8 Plenum discussions

Plenum discussions identified a number of relevant areas for which conclusions are summarized in the following.

The seminar discussions pointed out the need for increased support for operations and outreach for testbeds to amortize their investment. Testbeds are scientific instruments that allow scientists to deploy and measure innovative scientific phenomena. To fulfill this mission their capabilities are often complex, innovative, and need to evolve rapidly to follow the frontier of science. We note that the rapid rate of evolution means that testbed operations require relatively more effort than production resources that provide well-defined and understood capabilities. This includes a relatively high development effort to keep pace with requirements for new capabilities, and relatively high operational effort.

Testbeds also require relatively more interface between users and the infrastructure because they target specific needs; often offer unique capabilities; and support high degree of customization (i.e., "wide" interfaces). This is in contrast with production systems most of which have a relatively narrow set of interfaces (e.g., job submission via batch schedulers) that is also widely adopted so that its potential for problem solving is well understood. The

uniqueness of testbed capabilities also means that those capabilities are relatively less well known and their variety means that mapping the right problems to the right testbeds can be challenging.

Reproducibility is a key element of scientific method – whether deployed from the perspective of validating (and potentially sharpening) results or for the purpose of directly building on the results of others – in either case it fosters better understanding of results and promotes scientific debate between the author and the reproducer or reviewer of experiments that leads to sharper and riches scientific exchange.

Testbeds play a key role in the support for reproducibility, especially in computer science experiments which often rely on access to unique architectures and configurations, which together form an experimental environment. For many systems experiments, unless both the author and reproducer have access to the same experimental environment, the experiments frequently cannot be reproduced. The availability of open testbeds, available to both authors and reproducers, is therefore a critical ingredient of sustaining scientific reproducibility in computer science research. Providing this critical ingredient also means that testbeds provide an opportunity to promote the practice of reproducibility among the communities clustered around them. This reproducibility can be promoted via investment in developing methods for better packaging of experiments, repositories of existing artifacts, development of curricula based on reproducibility, and work with the community to contribute and adopt reproducibility practices in their work on both the "suppy" and the "demand side".

The potential of a malleable experiment workflow toolchain to increase the value of research testbeds, and the scientific productivity of the computer networks community has been identified. Compared to the relatively large amount of funding required for the large testbed research infrastructures, it appears attractive to initiate efforts to guide the further developments of the testbed community into a direction that directly meets the requirements of the users of the testbeds. With tight resources, it is necessary to prioritize these developments to increase the efficiency of the development process and maximize the impact it has on research. It is essential to initiate the dialog between these communities to reach these goals.

In the final plenum discussions, the seminar agreed on a set of three main conclusions, and 12 key recommendations, which are listed in the Executive Summary (cf. Section 1) of this report.

Participants

Tom Barbette UCLouvain, BE Terry Benzel $\rm USC\text{-}I\tilde{S}I$ – Marina del Rey, US Georg Carle TU München, DE Hakima Chaouchi IMT - Palaiseau, FRWalid Dabbous INRIA – Sophia Antipolis, FR Yuri Demchenko University of Amsterdam, NL Serge Fdida Sorbonne University – Paris, FR Sebastian Gallenmüller TU München – Garching, DE Jorge Gasos European Commission -Brussels, BE Michael Goedicke Universität Duisburg – Essen, DE

Cheikh Ahmadou Bamba Gueye Université Cheikh Anta Diop de Dakar, SN Tobias Hoßfeld Universität Würzburg, DE Kate Keahey Argonne National Laboratory & University of Chicago, US Wolfgang Kellerer TU München, DE Raymond Knopp EURECOM - Biot, FR Jim Kurose University of Massachusetts Amherst, US Deep Medhi NSF – Alexandria, US Jelena Mirkovic USC-ISI – Marina del Rey, US Andrew W. Moore University of Cambridge, GB

Paul Michael Ruth RENCI – Chapel Hill, US Damien Saucez INRIA – Sophia Antipolis, FR Björn Scheuermann TU Darmstadt, DE Henning Schulzrinne Columbia University -New York, US Jörg Widmer IMDEA Networks Institute -Madrid, ES Walter Willinger Niksun – Princeton, US Adam Wolisz TU Berlin, DE Ellen Zegura NSF - Alexandria, US Martina Zitterbart KIT – Karlsruher Institut für Technologie, DE

