CONTENTS

6 Executive Summary
10 Organization
12 ZIB Structure
14 ZIB Members
16 A New AI Department
18 Lise
20 Economic Situation in 2019
26 Green Energy
34 Riemannian Analysis of Time-Varying Shape
Data - Understanding Geometric Evolutions
44 On the Road to Autonomous Operating Room Scheduling
52 Solving Real-World Optimization Problems
66 Machine Learning and Big Data
76 Why Is There a New Wave of Attention for FPGAs?
86 References
89 Publications
98 Imprint
EXECUTIVE SUMMARY

The year 2019 was a very intense year for ZIB, in which a series of long-planned and relentlessly pursued goals were achieved and a plethora of new things were started. Selected successes in random order:

1. We were able to welcome a new Vice President to ZIB and establish him and around him a new department AI in Science, Society, and Technology (AIS²T).
2. We created the new department in relation to a new Federal Competence Center for Big Data and Machine Learning in Berlin.
3. We started the activity of the new Cluster of Excellence MATH+ at ZIB.
4. We inaugurated our newest supercomputer.
5. We were part of the Berlin University Alliance’s successful proposal in the context of the German excellence strategy.
6. We could celebrate a significant increase in our core budget through the state of Berlin.

In the following, we will elaborate on some of the above to outline their respective importance for ZIB.

New Vice President. In September 2019, Professor Sebastian Pokutta started as ZIB’s new Vice President (VP). This ended a period of almost four years in which ZIB had to operate with a vacant VP position. We are therefore particularly happy that we were able to successfully fill this important position. Prior to joining ZIB and TU Berlin, Professor Pokutta, a mathematician by training, was the David M. McKenney Family associate professor at the School of Industrial and Systems Engineering and an associate director of the Machine Learning @ GT Center at the Georgia Institute of Technology. Professor Pokutta received the David M. McKenney Family Early Career Professorship in 2016, an NSF CAREER Award in 2015, the Coca-Cola Early Career Professorship in 2014, the outstanding thesis award of the University of Duisburg-Essen in 2006, as well as various best paper awards. His research is situated at the intersection of artificial intelligence and optimization.

BIFOLD. Machine Learning (ML) (together with decision-making) and Big Data (BD) are the key pillars of Artificial Intelligence (AI). Research in ML and BD has been an integral part of research activities at ZIB for ten years and is becoming a cornerstone of our research strategy. To galvanize our activities in this area, we decided to create a new research department “AI in Society, Science, and Technology (AIS²T)” at ZIB in fall 2019. The department is strongly integrated with already-existing departments. Additionally, a new Berlin-based Federal Competence Center, the “Berlin Institute for the Foundations of Learning and Data” (BIFOLD), was founded that integrates the former Berlin Center for Machine Learning (BZML) and the Berlin Big Data Center (BBDC) into a new structure that is funded by the German Federal Ministry of Education and Research (BMBF). ZIB, having been part of both BZML and BBDC, is part of the consortium supporting BIFOLD and will establish new research groups in this context. We will further discuss our current research activities as well as our strategy to strengthen and intensify our presence in this area in the feature article “Machine Learning and Big Data.”
**MATH+**. The Berlin-based Cluster of Excellence MATH+ succeeded in the Excellence Strategy research competition run by the German federal and state governments. MATH+ started in January 2019 as a joint project between the universities FU Berlin, HU Berlin, and TU Berlin as well as the two research centers WIAS and ZIB. The cluster is a cross-institutional and transdisciplinary Cluster of Excellence where researchers explore and further develop new approaches in application-oriented mathematics. A particular focus is merging mathematical modeling with mathematical techniques for analyzing ever-growing amounts of data in life and material sciences in energy and network research, and in the humanities and social sciences. The aim is to boost not only scientific progress, but also technological innovation and a comprehensive understanding of social processes. ZIB’s research strengths in application-oriented mathematics and data-driven research in a transdisciplinary context including translation to industry and society at-large is recognized as one of the cornerstones of MATH+. At the institute, MATH+ started with ten new research projects funded through the Cluster of Excellence. These projects include about 40 researchers at ZIB alone and we provide an inside perspective into the research conducted in two of those projects in the feature articles “Riemannian Analysis of Time-Varying Shape Data – Understanding Geometric Evolution” and “Green Energy.”

**HLRN.** On December 6, 2019, the newest supercomputer at ZIB was officially inaugurated. This fourth generation of supercomputer systems of the Northern German Supercomputing Alliance (HLRN) holds rank 40 in the latest TOP500 list of the fastest supercomputers in the world. It provides about six times more compute power and twice the online storage space than its predecessor, while consuming only 50% more electric power and occupying the same floor space. With its almost quarter of a million compute cores, this massively parallel system architecture provides a new level of parallelism to support computationally demanding scientific workloads from a broad spectrum of use cases. More details can be found in the article “The 40th-Fastest Supercomputer in the World.”

**Berlin University Alliance.** In July 2019, the Berlin University Alliance (BUA) of FU Berlin, HU Berlin, and TU Berlin, together with their university hospital Charité, were admitted into the Universities of Excellence funding line of the German government’s Excellence Strategy. The Berlin University Alliance’s long-term goal is to turn Berlin into an integrated research environment. ZIB is one of the BUA’s partner institutes for application-oriented mathematics and research infrastructure-related community building. The aim is to provide better access to and joint usage of top-level research infrastructure and community-centric services in order to foster interdisciplinary research communities around cutting-edge research infrastructure such as the HPC and large data facilities at ZIB.
Increase of budget. In mid-December 2019, as a much-celebrated Christmas gift, Berlin’s parliament approved an increase of almost 15% in the basic budget of ZIB for the years 2020 and 2021 in order to support ZIB’s current research activities, allowing for additional investments in restructuring and strengthening research, and improving research infrastructure. This highly necessary increase of our core budget and the complementing record high level of third-party funds bring ZIB into an economically strong position to tackle major research and innovation challenges in the future.

More insights. In addition to the topics already mentioned, this annual report provides insights into a variety of other success stories and gives a general overview of ZIB’s organization and key factors for its successful development. For example, the feature article “On the Road to Autonomous Operating Room Scheduling” provides insight into how algorithmic intelligence and optimal decision-making can be utilized in the operating room through a joint project between ZIB and several partners including Berlin’s university hospital Charité. This overview is complemented by the feature article “Solving Real-World Optimization Problems,” that reports on ZIB’s renowned in-house optimization suite SCIP and its success stories, outlining the possibilities of one of the fastest academic mathematical optimization software packages worldwide. In another feature article, “Reconfigurable Computing Today,” it is explained why there is a new wave of interest in Field-Programmable Gate Arrays (FPGA). FPGAs can be used for implementing highly data-parallel architectures to target more power-efficient solutions as well as accelerators for HPC.

Last year was a fantastic year for ZIB with key successes that will positively impact on ZIB’s development for a decade and beyond. Many other very positive developments make us confident that our institute has a bright future. Put differently, ZIB continues to be a place for excellent research and first-rate scientific services and infrastructure.
Zuse Institute Berlin -
The Movie

www.zib.de/movie  Click on the image to start the movie.
The Statutes

The Statutes, adopted by the Board of Directors at its meeting on June 30, 2005, define the functions and procedures of ZIB’s bodies, determine ZIB’s research and development mission and its service tasks, and frame upon the composition of the Scientific Advisory Board and its role.

Administrative Bodies

The bodies of ZIB are the President and the Board of Directors (Verwaltungsrat).

President of ZIB
Prof. Dr. Christof Schütte

Vice President
until February 28, 2019
Prof. Dr. Martin Skutella
since September 1, 2019
Prof. Dr. Sebastian Pokutta

The Board of Directors was composed in 2019 as follows:

Prof. Dr. Peter Frensch
Vice President, Humboldt-Universität zu Berlin (Chairman)

Prof. Dr. Christian Thomsen
President, Technische Universität Berlin (Vice Chairman)

Dr.-Ing. Andrea Bör
Provost, Freie Universität Berlin

Prof. Dr. Günther Ziegler
President, Freie Universität Berlin

Frau Ellen Fröhlich
Der Regierende Bürgermeister von Berlin, Senatskanzlei – Wissenschaft und Forschung

Dr. Jürgen Varnhorn
Senatsverwaltung für Wirtschaft, Energie und Betriebe

Prof. Dr. Manfred Hennecke
Bundesanstalt für Materialforschung und -prüfung (BAM)

Thomas Frederking
Helmholtz-Zentrum Berlin für Materialien und Energie (HZB)

Prof. Dr. Heike Graßmann
Max-Delbrück-Centrum für Molekulare Medizin (MDC)

The Board of Directors met on June 7, 2019, and November 27, 2019.
The Scientific Advisory Board advises ZIB on scientific and technical issues, supports ZIB’s work, and facilitates ZIB’s cooperation and partnership with universities, research institutions, and industry.

The Board of Directors appointed the following members to the Scientific Advisory Board:

Prof. Dr. Jörg-Rüdiger Sack
Carleton University, Ottawa, Canada

Prof. Dr. Cecilia Clementi
Rice University, Houston, Texas, USA

Prof. Dr. Michael Dellnitz
Universität Paderborn, Germany

Prof. Dr. Rolf Krause
Université della svizzera italiana, Lugano, Switzerland

Ludger D. Sax
Grid Optimization Europe GmbH

Prof. Dr. Reinhard Schneider
Université du Luxembourg, Luxembourg

Prof. Dr. Dorothea Wagner
Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany

The Scientific Advisory Board met on July 8 and 9, 2019, at ZIB.
ZIB is structured into four divisions: three scientific divisions and ZIB’s administration.

Each of the scientific divisions is composed of two departments that are further subdivided into research groups (darker bluish color) and research service groups (lighter bluish color).

**LEGEND**

- Scientific divisions and departments
- Research groups
- Research Service groups
- Core Facility
ZIB MEMBERS

In fall 2019, Professor Sebastian Pokutta from TU Berlin was appointed Vice President of the Zuse Institute Berlin. Having received both his diploma and Ph.D. in mathematics from the University of Duisburg-Essen in Germany, Professor Pokutta was a postdoctoral researcher and visiting lecturer at MIT, worked for IBM ILOG, and Krall Demmel Baumgarten. Prior to joining ZIB and TU Berlin, he was the David M. McKenney Family associate professor in the School of Industrial and Systems Engineering and an associate director of the Machine Learning @ GT Center at the Georgia Institute of Technology as well as a professor at the University of Erlangen-Nürnberg. Professor Pokutta received the David M. McKenney Family Early Career Professorship in 2016, an NSF CAREER Award in 2015, the Coca-Cola Early Career Professorship in 2014, the outstanding thesis award of the University of Duisburg-Essen in 2006, as well as various best paper awards. His research is situated at the intersection of artificial intelligence and optimization. A particular focus is on combining machine learning with optimization techniques, both discrete and continuous.

With the appointment of the new Vice President, the Zuse Institute Berlin is further expanding its foray into artificial intelligence, machine learning, and optimization. Apart from the integration with Berlin’s AI and ML efforts, such as BIFOLD, a particular focus is on international collaborations and partnerships. To date, research collaborations and memorandums of understanding (MoU) have been signed with RIKEN-AIP in Tokyo, which is comparable to a Max Planck Institute, the McMaster University in Toronto, as well as the Fraunhofer IIS in Erlangen.
Tightly integrated with the activities of the VP is the newly founded department with the name AI in Society, Science, and Technology (AIST). “AI is not just a methodology but a paradigm shift that permeates all disciplines. ZIB is embracing the opportunities and possibilities of this technology. At the same time, we do understand that this technology is touching human lives and society on so many levels that an inquiry disregarding societal aspects would be irresponsible,” says Professor Pokutta. The department will focus on the development of new AI methodologies as well the interface and translation of these methods into an industry context. Initially, the department will be concerned with the following three major directions:

1. **Integrating Learning and Decision-Making.** We are at a point in time where we understand how to turn data into insights via machine learning very well. We also very much know how to compute optimal decisions using prior insights obtained, e.g., from data. However, the holistic integration of learning and decisions is still a key open problem and one major challenge that we aim to tackle.

2. **Specialized Hardware.** Today’s SoCs and FPGAs are powerful customizable hardware that allows complex algorithms to be accelerated and implemented for specialized loads. While specialized/customized hardware is quite common in the context of deep learning (e.g. by means of GPUs or Google’s TPU, which one might think of as custom ASICs), no such custom implementations exist for discrete methods (such as mixed-integer programming algorithms or constraint programming algorithms) or heuristics commonly used in discrete methods. We are interested in designing custom hardware accelerators (via FPGAs) to significantly speed things up, i.e. discrete optimization algorithms by implementing specialized operations/functions. We are also interested in the edge computing regime, where we want to deploy complex algorithms in situ. An example in this context is, for example, the deployment of SCIP on a Raspberry Pi.

3. **Artificial Intelligence in Society.** A particular focus is on the interaction of AI methods with societal challenges as it has become clear that questions of fairness, accountability, transparency, and explainability (FATE) have to be not only addressed through policies but also directly within the algorithmic design.
The new HLRN-IV supercomputer at ZIB is a massively parallel system for a broad spectrum of applications.

On December 6th, 2019, the next-generation supercomputer at ZIB was officially inaugurated. This fourth generation of supercomputer systems of the North German Supercomputing Alliance (HLRN) holds the 40th rank in the latest TOP500 list of the fastest supercomputers of the world.

With its name “Lise”, we honor the theoretical work of Lise Meitner to explain the experiments of Otto Hahn on nuclear fission of uranium in the late 1930s. Lise Meitner spent most of her scientific career in Berlin, where she was a physics professor and a department head at the Kaiser Wilhelm Institute (today Hahn-Meitner Institute), just a few hundred meters away from ZIB. Notable, she was the first woman to become a full professor of physics in Germany.

The HLRN-IV supercomputer “Lise” provides about six times more compute capacity and twice of the online storage space than its predecessor, the HLRN-III system “Konrad”, while consuming only 50% more electric power and occupying the same floor space. With its almost a quarter of million compute cores, this massively parallel system architecture provides a new level of parallelism to support computationally demanding scientific workloads out of a broad spectrum of usage scenarios.

The HLRN-IV System “Lise” at a Glance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Racks</td>
<td>14 with a mixed warm-water and air cooling infrastructure.</td>
</tr>
<tr>
<td>Fat Tree</td>
<td>is the interconnect topology of the Intel Omni-Path network providing a low latency of 1.65 µs maximum and a high bandwidth, i.e., 7.8 TB/s bisection bandwidth, for the communication across the entire system. This interconnect is realized with two 1152-port OPA100 director switches and 54× 48-port edge switches.</td>
</tr>
<tr>
<td>PCIe Lanes</td>
<td>8 PFlop/s</td>
</tr>
<tr>
<td>HPC nodes</td>
<td>1,270 in total, each with two Intel Xeon Cascade Lake Advanced Processors with 48 cores, i.e. 96 Intel Xeon cores per node: 1,236 standard nodes with 384 GByte main memory, 82 large memory nodes with 768 GByte memory, and 2 huge memory nodes with 1,536 GByte memory per node.</td>
</tr>
<tr>
<td>CPU Cores</td>
<td>121,920</td>
</tr>
<tr>
<td>CPUs</td>
<td>2,540 with 48 cores each, i.e. 96 Intel Xeon cores per node.</td>
</tr>
<tr>
<td>Memory</td>
<td>8.4 PByte</td>
</tr>
<tr>
<td>Distributed Main Memory</td>
<td>502 TByte</td>
</tr>
<tr>
<td>Online Storage Capacity</td>
<td>8.4 PByte</td>
</tr>
</tbody>
</table>

Introducing HLRN-IV
The HLRN-IV Opening Event

The official inauguration of the HLRN-IV supercomputer on December 6th was held at ZIB and was accompanied by high-profile guests and prominent speakers. Witnessed by state ministers and state secretaries of the seven HLRN federal states, Christof Schütte (ZIB), Wolf-Dieter Lukas (BMBF), Steffen Krach (Berlin), Björn Thümler (Lower Saxony), Ursula Morgenstern (Atos) and Hannes Schwaderer (Intel) delivered welcome greetings before pushing the red button to start the official operation of the new HLRN-IV system in Berlin. The highlight of the opening ceremony was the inspiring and entertaining keynote speech of Horst Simon, deputy director of the Berkeley Lab and co-author of the TOP500 list. In his keynote “Supercomputers and Superintelligence” he emphasized the tight relation of HPC and AI by illustrating the enormous potential and challenges of artificial intelligence solutions based on large-scale computations. Finally, in the afternoon session the impact of supercomputer resources for the scientific progress was emphasized. Joachim Sauer, Bettina Keller, and Siegfried Raasch presented their latest results on such diverse topics like heterogeneous catalysis, the dynamics of bio-molecular complexes, and urban climate simulations, and they explained why increasingly large supercomputer capacities are needed to advance their science domains.

Storage Infrastructure of the HLRN-IV “Lise”

- 8.4 PByte online storage capacity is available in a globally accessible parallel file systems (Lustre) offering a high bandwidth access to application data in batch jobs.
- 340 TByte online storage capacity is provided in a globally accessible NAS appliance for permanent project data and program code development.
- 110 PByte Peta-Scale tape library for long-term archiving of large data sets; operated independently by ZIB.
In 2019, the total income of ZIB comprised 25.1 million euros. The main part of this was made available by the Federal State of Berlin as the basic financial stock of ZIB (10.3 million euros) including investments and Berlin’s part of the budget of HLRN at ZIB. The second largest part of the budget resulted from third-party funds (8.6 million euros) acquired by ZIB from public funding agencies (mainly DFG and BMBF) and via industrial research projects. This was complemented by a variety of further grants, such as the HLRN budget made available by other German states or the research service budget of KOBV, summing up to almost 6.2 million euros in total.
The Zuse Institute Berlin (ZIB) finances its scientific work via three main sources: the basic financial stock of the Federal State of Berlin and third-party funds from public sponsors and those of industrial cooperation contracts.

In 2019, ZIB raised third-party funding through a large number of projects. Project-related public third-party funds declined from 6.425 million euros in 2018 to 6.137 million euros in 2019, while industrial third-party projects rose from 1.707 million euros to 2.450 million euros. In total, 8.587 million euros in third-party funding marked again a new record in ZIB’s history – an increase for the eighth year in a row.

**ZIB THIRD-PARTY FUNDS BY SOURCE**

- €2,450,000 (29% Industry)
- €2,216,450 (23% DFG incl. FC Modal)
- €2,017,370 (26% BMBF incl. FC Modal)
- €1,902,750 (22% Other Public Funds)
ZIB THIRD-PARTY FUNDS IN EUROS

2008 2009 2010 2011 2012 2013

€2,000,000
€4,000,000
€6,000,000
€8,000,000
€10,000,000
### Spin-Offs

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Year</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RISK-CONSULTING</td>
<td>1994</td>
<td><a href="http://www.risk-consulting.de">www.risk-consulting.de</a></td>
<td>Database marketing for insurance companies</td>
</tr>
<tr>
<td>Intranetz GmbH</td>
<td>1996</td>
<td><a href="http://www.intranetz.de">www.intranetz.de</a></td>
<td>Software development for logistics, database publishing, and e-government</td>
</tr>
<tr>
<td>AktuarData GmbH</td>
<td>1998</td>
<td><a href="http://www.aktuardata.de">www.aktuardata.de</a></td>
<td>Development and distribution of risk-evaluation systems in health insurance</td>
</tr>
<tr>
<td>Visage Imaging GmbH</td>
<td>1999</td>
<td><a href="http://www.visageimaging.com">www.visageimaging.com</a></td>
<td>Advanced visualization solutions for diagnostic imaging</td>
</tr>
<tr>
<td>atesio GmbH</td>
<td>2000</td>
<td><a href="http://www.atesio.de">www.atesio.de</a></td>
<td>Development of software and consulting for planning, configuration, and optimization of telecommunication networks</td>
</tr>
<tr>
<td>bit-side GmbH</td>
<td>2000</td>
<td></td>
<td>Telecommunication applications and visualization</td>
</tr>
<tr>
<td>JCMwave GmbH</td>
<td>2005</td>
<td><a href="http://www.jcmwave.com">www.jcmwave.com</a></td>
<td>Simulation software for optical components</td>
</tr>
<tr>
<td>onScale solutions GmbH</td>
<td>2006</td>
<td><a href="http://www.onscale.de">www.onscale.de</a></td>
<td>Software development, consulting, and services for parallel and distributed storage and computing systems</td>
</tr>
<tr>
<td>Laubwerk GmbH</td>
<td>2009</td>
<td><a href="http://www.laubwerk.com">www.laubwerk.com</a></td>
<td>Construction of digital plant models</td>
</tr>
<tr>
<td>1000shapes GmbH</td>
<td>2010</td>
<td><a href="http://www.1000shapes.com">www.1000shapes.com</a></td>
<td>Statistical shape analysis</td>
</tr>
<tr>
<td>TASK – Berthold Gleixner Heinz Koch GbR</td>
<td>2010</td>
<td></td>
<td>Distribution, services, and consulting for ZIB’s optimization suite</td>
</tr>
<tr>
<td>Quobyte Inc.</td>
<td>2013</td>
<td><a href="http://www.quobyte.com">www.quobyte.com</a></td>
<td>Quobyte develops carrier-grade storage software that runs on off-the-shelf hardware</td>
</tr>
<tr>
<td>Keylight GmbH</td>
<td>2015</td>
<td><a href="http://www.keylight.de">www.keylight.de</a></td>
<td>Keylight develops scalable real-time Web services and intuitive apps. The focus is on proximity, marketing, iBeacon, and Eddystone for interactive business models</td>
</tr>
<tr>
<td>DoloPharm Biosciences UG</td>
<td>2017</td>
<td></td>
<td>A specialty pharmaceutical company focused on the clinical and commercial development of new products in pain management that meet the needs of acute and chronic care practitioners and their patients</td>
</tr>
<tr>
<td>Keopsion GmbH</td>
<td>2018</td>
<td><a href="http://www.keopsion.com">www.keopsion.com</a></td>
<td></td>
</tr>
</tbody>
</table>
In 2019, 199 people were employed at ZIB; of these, 119 positions were financed by third-party funds. The number of employees decreased in comparison to 2018, mainly because of a funding gap between the first and second funding periods of the Research Campus MODAL.

<table>
<thead>
<tr>
<th></th>
<th>1/1/2019</th>
<th>1/1/2020</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Permanent</td>
<td>Temporary</td>
</tr>
<tr>
<td>MANAGEMENT</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>SCIENTISTS</td>
<td>22</td>
<td>96</td>
</tr>
<tr>
<td>SERVICE PERSONNEL</td>
<td>35</td>
<td>6</td>
</tr>
<tr>
<td>KOBV HEADQUARTERS</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>STUDENTS</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>152</td>
</tr>
</tbody>
</table>

* without temporary management
Artistic view of nanoparticles for energy conversion on top of a metasurface allowing the conversion efficiency to be enhanced through an excited photonic resonance [1].
Renewable energy generation using photovoltaics, photocatalysis, photosynthesis, or other methods can help to reduce the CO$_2$ footprint of modern society. Hereby, solar energy is converted to other exploitable forms of energy, like electrical or chemical energy. For a better understanding and an effective design of the physical processes and of next-generation devices, advanced modeling and simulation methods are essential.
Energy Research at ZIB

Several of ZIB’s cooperations are addressing various topics in the field of green energy. Within the Cluster of Excellence MATH+, a cooperation between Helmholtz Zentrum Berlin für Materialien und Energie (HZB) and ZIB, investigates numerical methods needed to design optical energy conversion processes. Within the Helmholtz Excellence Network SolarMath, HZB and ZIB jointly investigate perovskite-silicon tandem solar cells and hybrid components for solar fuel manipulation. The training and networking of young scientists is supported within the Berlin Joint Lab for Optical Simulations in Energy Research (BerOSE) as well as in the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS).

Solar Fuel Devices Can Convert Sunlight Directly into Long-Term Storable Fuels

The energy supply network needs to go through tremendous change in order to meet the demands to mitigate climate change. To achieve this, every aspect of the supply chain must be carefully evaluated and optimized. A cornerstone of any sustainable energy network will be solar energy. While conventional solar cells are rapidly increasing their penetration into energy markets, the requirement of electricity networks to provide on-demand power is at odds with the intermittent supply of solar cells. Solar fuel devices are able to convert solar energy directly into usable fuels, thus storing the energy and allowing distribution through existent networks. Of the various fuels that can be created using photochemistry, hydrogen, obtained from water splitting, is the most prominent candidate due to the abundant, low-cost supply of water. Therefore, a better understanding of solar-driven water splitting and developing the technology for water splitting on a large scale is a central challenge facing the transition to a green economy. This technology will likely first be based on the currently available photovoltaic (PV) powered electrolysis. However, it is expected that the next technology step will be to integrate light absorption and electrochemistry into a single device. Optical device design is an important problem for both of these technology steps: one of the limitations of both currently available PV technology and the developing solar fuel technologies is insufficient absorption of light due to reflective losses, transmissive losses, and parasitic absorption. In order to overcome these challenges, modeling, simulation, and optimization [2, 3] plays a crucial role. Proper light management in many...
photovoltaic devices has already lead to large increases in efficiency, typically through the use of nanostructures. In a collaboration between Helmholtz-Zentrum Berlin für Materialien und Energie and Zuse Institute Berlin, we have recently investigated concepts for absorption enhancement in integrated solar fuel devices. It could be shown that nanostructuring of the device can significantly increase efficiency, and that plasmonic resonances in the optical excitation of nanoparticles can play a major role in optimized devices [5].

Fig. 1: Artistic view of a solar fuel device. Light is absorbed in a semiconductor material, CuBi$_2$O$_4$, generated electrons and holes drive the splitting of water to hydrogen (H$_2$) and oxygen (O$_2$, not shown). Metal (silver, Ag) nanostructures enhance the efficiency through providing plasmonic resonances (inset with electromagnetic field distribution in pseudocolor scale).
Photosynthesis fuels life on earth by converting incoming solar radiation into chemical energy. Some photosynthetic organisms have adapted particularly well to low light environments. For instance, green sulfur bacteria use “leftover” near-infrared light to thrive at places where other wavelengths have already been absorbed by other species. The sulfur-rich volcanic ponds in the Yellow Stone National Park provide a suitable natural habitat.

To further concentrate the available light, green sulfur bacteria developed a multi-stage photon trap: starting from a large antenna complex, grounded on a base-plate structure, the energy is directed toward the reaction center [5]. This link is established by the Fenna-Matthews-Olson complex, the first photosynthetic complex whose structure was determined by X-ray crystallography [6]. The bacteriochlorophyll-a complex embedded in the photosynthetic apparatus is tuned by surrounding proteins to establish an energetic funnel that guides the photoexcitation toward the reaction center. The directed energy transfer comes at a price: part of the light energy is converted to heat in order to guide the larger remaining part to the reaction center.

Sunlight is composed of light with different wavelengths. Green sulfur bacteria are harvesting near-infrared photons (750 nm), whereas plants absorb light in the blue and red colors, giving rise to the green color of leaves.
Since 2005, it has been possible to directly image the energy transfer using a sequence of laser pulses that first excite and later – at a specific time delay – probe the energetic state of the molecular complex [7]. The energy transfer is then directly monitored by the movement of emission peaks. At ZIB, we have developed methods to efficiently compute and predict the energetic pathways and to elucidate the intermediate steps. To capture the crucial role of the heat transferred to molecular vibrations, these computations make use of the quantum-mechanical density matrix description [8, 9]. The strong interaction of vibrational and electronic excitations forms an entangled state of the molecular complex.

Recent experiments employ fast-varying polarization changes of laser pulses to further probe the configuration and arrangement of the bacteriochlorophylls. The modeling of the signals requires considering large ensembles of molecular complexes, a task ideally suited for the HLRN supercomputer [12]. The resulting large data sets are efficiently encoded in neural networks using machine learning and are used to investigate the arrangement of different photosynthetic complexes.

Schematics of the photosynthetic apparatus of green sulfur bacteria (C. tepidum), adapted from [5].

Time sequence of the energy transfer in green sulfur bacteria (C. tepidum). The horizontal axis denotes the excitation energy and the vertical axis the de-excitation energy. A decay of intensity to the lower diagonal tracks the energy transfer [10, 11].
Light Management for Highly Efficient Photovoltaic Devices

Nanotextured Perovskite Solar Cells

Currently, perovskite-silicon tandem solar cells are the most investigated concept to overcome the theoretical limit for the power conversion efficiency of single-junction silicon solar cells. Optical simulations are extremely valuable to study the distribution of light within the solar cells, and allow the minimization of losses from reflection and parasitic absorption. For monolithic perovskite-silicon solar cells, it is vital that the available light is equally distributed between the two subcells, which is known as current matching. At ZIB, we develop advanced finite-element-method-based (FEM) simulation tools for optimizing the solar-cell stacks and nanotextures. These tools are used to study, for example, how different light management approaches influence the sensitivity of the solar module to the illumination condition [13].

Tetrahedral mesh for a FEM simulation of electromagnetic field absorption in a nano-structured perovskite-silicon tandem solar cell. Colors indicate different materials (green: perovskite, blue: silicon).
Realistic Weather Conditions Impact
Bifacial Solar-Cell Installations

Bifacial solar cells are a promising technology with a quickly increasing market share. By converting light received at both sides of the module, the power output of a solar cell can be significantly increased. Modern cell designs are intrinsically suitable for bifacial modules and therefore the difference in production costs of bifacial compared to classic modules is relatively small. It is however challenging to estimate the energy output and to find optimal configurations of bifacial photovoltaic installations. Within the Joint Lab BerOSE, models for data-driven optimization of modules in field installations have been developed. This allows the amount of light that falls on the front- and backside of a module to be computed. The tilt angle of the modules and the distance between the module rows are important parameters that determine how much electricity can be generated. With increasing distance between the rows, the power per module will increase, however it also becomes more cost intensive due to the increasing ground consumption. In order to find the optimal configurations of tilt and row distance for realistic weather data, the combination of an economical model and a tool for optical simulation has been used. This allows the levelized cost of electricity (LCOE) that relates the total amount of energy produced by a power plant to the total cost over its lifetime to be calculated. The optical simulation can be performed with weather data that is readily available for large parts of the world. By using Bayesian optimization, we obtain parameters that allow LCOEs that are significantly reduced [14], compared to rules of thumb, as they are employed in current state-of-the-art installations.

Model of an illuminated installation of bifacial solar-cell modules. Depending on the tilt angle of the modules and on the module spacing, different portions of direct sunlight and of sunlight scattered from the ground reach the front and back surfaces of the solar cell. Right: Levelized cost of electricity (LCOE) depending on module tilt and spacing, for realistic weather data and economic model at a specific location of the installation in the US. The optimum installation parameters allow for significantly reduced costs with respect to rule-of-thumb optimization.
RIEMANNIAN ANALYSIS OF TIME-VARYING SHAPE DATA - UNDERSTANDING GEOMETRIC EVOLUTIONS

Evolution of mitral valve shape during distole phase of the cardiac cycle.
Material objects, both in the inanimate and the animate world, are characterized by their material and their shape. In our work we focus on the geometric aspect, i.e. the shape. Instead of characterizing shapes by a few meaningful parameters, we consider shapes in their entirety. This allows a full geometric characterization of empirically given shapes and more differentiated assertions about these. From a mathematical point of view, we consider individual shapes as single points in a high-dimensional space, the so-called “shape space”.

For many applications, e.g. in medicine or biology, the characterization of sets of shapes is of interest: What types of shapes do occur in a given ensemble? How frequently do the different types of shape show up? What is a typical and what is an atypical shape? These are questions where statistics comes into play, or mathematically speaking, statistics in high-dimensional shape spaces. In recent years, appropriate mathematical theories and computational tools have been developed. Statistical shape modeling makes it possible to consider all shape features and their correlations at once without having to pre-define discrete shape measurements. The complete coverage of the shape, including the correlations of the degrees of freedom, allows more differentiated conclusions and provides better quantitative measures that are statistically significant.

The focus of our project is the analysis of time-varying shapes. These exist in abundance, both in everyday life and in science – especially in life, environmental and earth sciences, but also in engineering and cultural sciences. A large class in life sciences are biological changes of shapes within and between individuals. Such changes can be tracked over time in longitudinal imaging studies, e.g. to gain insights into dynamic processes, such as ageing or disease progression. Another example is tracking the beating heart during the cardiac cycle, where atypical shape changes can indicate certain diseases and precise analysis of shape changes can potentially provide further diagnostic information. Also quite different time scales can be involved: for example in the analysis of the stylistic development of ornaments over cultural periods, or the analysis of the evolutionary development of bony structures in living beings.

We have extended the mathematical tools for the statistical analysis of (static) shapes to the analysis of time-varying shapes – or in mathematical terms, to shape trajectories, i.e. curves in the shape space. These trajectories are discretely sampled by the empirically given shapes. They are themselves geometrical objects and thus amenable to geometric techniques. Since each individual shape is defined by many degrees of freedom and for statistical reasons many objects have to be included in the analysis at many points in time, we are dealing with “big data”. A central concern is therefore the development of efficient algorithms.
For a mathematical concept of "shape", we can start directly from the everyday concept in which we refer to the outer boundary (i.e. the surface) of an object. A possible framework for the formalization of the concept of shape is based on the comparison of related forms: An object class under study can be represented by a common deformable template that accounts for the typicality of the objects' structure. The shape variability is represented by deformations that are applied to the template. Codifying shapes in such a way allows to interpret them as elements in a high-dimensional space of deformations. This so-called configuration space not only encodes the geometric form of objects but also their scale, position and orientation within the 3D space they are embedded in. By identifying shapes through similarity transformations we obtain the shape space [3]. This is suitable to statistical shape analysis. The last step (mathematically, a quotient taking through a group action), however, introduces curvature to shape space. Contrary to flat spaces, shortest connecting paths in shape space are not straight lines but curved trajectories referred to as geodesics (see Fig. 1).

Fig. 1: Visualization of shortest paths, i.e. geodesics, connecting two body shapes w.r.t. the flat ambient space (red) and a curved shape space (green). The latter contains only valid shape instances whereas the former contains shapes with artifacts, like shrunken arms.
The nonlinearity of shape space further implies that there is no global system of coordinates. Consequently, there is no direct way to compare pairwise differences between shapes. To obtain a consistent description of structural changes at the population level (e.g. for group-wise analysis), the differences need to be transferred into a common reference frame. Among the different techniques proposed, constructions based on parallel transport provide the most natural approach. As parallel transport is rarely given in closed form, in general it has to be approximated numerically, e.g. employing Schild’s ladder or fanning. In particular, except for the limited setting of planar shapes, this is the case for Kendall’s shape space [5]. Utilizing closed form expressions of Kendall’s pre-shape sphere, we reduce parallel transport to the solution of a homogeneous first-order differential equation that allows for highly efficient computations [7]. Moreover, we reduce the important case of parallel transport along a geodesic path to the solution of a low-dimensional equation that only depends on the dimension of the ambient space and not on the spatial resolution of the discrete representation.

Whereas the nonlinearity of the shape space ensures consistency, e.g. by preventing bias due to misalignment of shape observations, it also impedes the application of classical statistical tools. As a fully intrinsic treatment of the analysis problem can be computationally demanding, a common approach is to approximate it using extrinsic distances. For data with a large spread in shape space or within regions of high curvature, such linearization will introduce distortions that degrade the statistical power [9]. Therefore, we derive novel geometric structures that allow for efficient shape analysis and, thus, facilitate applications that require interactive response or involve large shape populations. To this end, we employ differential coordinates that are derived from the (deformation) gradient of the map that encodes the shape relative to the template and, hence, naturally belong to the group of orientation preserving linear transformations \( \text{GL}^+(3) \) [2]. Performing intrinsic calculus on this representation allows for fast computations while, at the same time, accounts for the nonlinearity in shape variation. Following a surface-theoretic approach, we further extend the differential coordinates based on discrete fundamental forms to derive a shape representation that is invariant under Euclidean motion and thus alignment-free [1]. We endow this representation with a Lie group structure that admits bi-invariant metrics and therefore allows for consistent analysis using manifold-valued statistics based on the Riemannian framework. The rich structure of the derived shape space yields highly discriminative shape descriptors providing a compact representation that is amenable to learning algorithms. We evaluate the performance of our model w.r.t. shape-based classification of pathological malformations of the human knee and show that it outperforms state-of-the-art approaches especially in the presence of sparse training data.
Regression

Given empirically-defined shape trajectories we would like to estimate the relationship between the observed variables, i.e. the shapes and their co-varying parameters. To this end, regression analysis is a reliable statistical approach to identify which variables have an impact on a quantity of interest. The process of performing regression allows one to confidently determine which parameters matter most, which parameters can be ignored, and how these parameters influence each other. Typically, the evolution of shapes is assumed to be smooth and related to a single explanatory variable (usually time), for example when studying growth patterns. In this context, spatiotemporal regression models allow to estimate continuous trajectories from sparse and potentially noisy samples. They also provide a way to describe the data at unobserved times (i.e. shape changes between observation times and — within certain limits — also at future times). Such a time interpolation of data further allows to compare shape evolutions between different subtypes that have been observed at unequal time points.

As shape spaces lack a global vector space structure, any linear combination of shapes may not lie on the manifold. Furthermore, embedding the manifold-valued variables in a Euclidean space might introduce distortions and, thus, results in a poor estimation of the model. These problems advocate the development of novel, intrinsic regression methods for non-Euclidean data. The most widely used approach is to approximate the observed temporal shape data by geodesics in shape space. Geodesic models are attractive as they feature a compact representation (similar to the slope and intercept term in linear regression) and therefore allow for computationally efficient inference. In particular, the goal of geodesic regression is to find a geodesic curve in shape space that best fits the data in a least-squares sense.

In the absence of an analytic solution, the regression problem has to be solved numerically. Again, to obtain consistent and efficient computational schemes the non-Euclidean structure has to be taken into account. To this end, the gradient of the cost function can be computed using Jacobi fields (Fig. 2, right), since they express the derivatives of the exponential map. Based on an analytic derivation for these expressions [7], we can fully leverage the geometry of the shape space using Riemannian optimization procedures. Such an intrinsic approach leads to highly efficient algorithms reducing the computational expense by several orders of magnitude over general, nonlinear constrained optimization.

Fig. 2: Schematic depiction of best fitting geodesic (left) and a Jacobi field (right).
Hierarchical Models

Processes such as disease recovery, style progression of prehistoric remains, or growth are inherently time-dependent, requiring measurements at multiple time points to be sufficiently described. To gain insight into such dynamical processes, morphological studies rely on longitudinal datasets that capture shape changes within and across subjects over time. When analyzing such observations of shape trajectories we have to distinguish between morphological differences due to (i) temporal shape evolutions of a single subject and (ii) the geometric variability in a population (see Fig. 3 for an illustration). While approaches for the analysis of time series of scalar data are well understood and routinely employed in statistics and medical imaging communities, generalization to complex data such as shapes are at an early stage of research. Methods designed for cross-sectional data analysis, e.g. regression, do not consider the inherent correlation of repeated measurements of the same subject, nor do they inform how an instance relates to a population-average trend. Therefore, analysis methods for longitudinal datasets must capture and disentangle the cross-sectional variability in shape and the temporal variability due to underlying processes of change. Fig. 4 shows a synthetic example where cross-sectional regression fails to estimate a population-average trend correctly.
These problems motivate the use of hierarchical models that include intra-individual changes in the response variable and thereby have the ability to differentiate between cohort and temporal effects. In the first stage of a hierarchical model, inner-individual changes are modeled as smooth, parametric curves determined via spatiotemporal regression. In the second stage, these subject-specific trends are considered as disturbances of a population-average trajectory. To this end, a principled way of comparing shape trends is needed. In particular, this requires a notion of distance for shape trajectories that is consistent with the Riemannian metric of the underlying shape space. As for manifold-valued regression, state-of-the-art approaches model shape trends as geodesic trajectories, which can be parametrized as points in the tangent bundle of shape space [6]. While the Sasaki metric is a natural metric on the tangent bundle, its geodesic computations require time-discrete approximation schemes involving the Riemannian curvature tensor. This not only incurs high computational costs but also impacts numerical stability. We consider a novel approach that overcomes these shortcomings [8]. To this end, we identify tangent vectors of the tangent bundle with vector fields along the geodesic trend. This provides a notion of a canonical metric that is motivated from a functional view of parameterized curves in the shape space. Considering the space of the geodesics as a submanifold in the space of shape trajectories, this allows in particular the use of a naturally induced distance. The corresponding shortest path, logarithmic map and average geodesic, can be computed by variational time-discretization. Remarkably, the underlying energy function allows for fast and simple evaluation, increasing computational efficiency. In particular, it neither requires curvature computation nor decomposition in horizontal and vertical components.

Based on the derived metric for geodesic trends, we obtain a notion of mean, covariance, and Mahalanobis distance. This allows us to perform a statistical hypothesis test for comparing group-wise mean trends. More precisely, we can test for significant differences in average shape trajectories using a manifold-valued Hotelling $t^2$ statistic in a non-parametric permutation test setup. This framework allowed us to reveal differences between physio- and pathological shape evolution in a long-term study on incident knee Osteoarthritis [8].

Fig. 4: Sketch of longitudinal dataset together with population-average trends estimated by cross-sectional (red) and a hierarchical (green) model.
Current and Future Work

The many advances we have made in recent years in the analysis of shapes and shape trajectories open up new avenues and allow us to address remaining challenges to further extend the scope of the derived methodologies.

Higher-Order Models

Non-monotonous shape changes, e.g. present in time-series of cardiac shape motion or anatomical changes in the human brain over the course of decades, do generally not adhere to constraints of geodesicity. This necessitates the development of higher-order models. To this end, we are generalizing our approaches to Riemannian spline models based on constructive algorithms [4]. Fig. 5 shows a first preliminary result of a spline-based regression of a mitral valve shape trajectory.
Recent advances in the field of machine learning have led to qualitative breakthroughs on a wide variety of tasks. Consequently there is a high interest of translating learning approaches to the high-dimensional and non-Euclidean setting of shapes and time-series thereof. In [10] we derived a transductive learning approach for morphometric osteophyte grading based on geometric deep learning (see Fig. 6). We formulate the grading task as semi-supervised node classification problem on a graph embedded in shape space. To account for the high-dimensionality and nonlinear structure of shape space we employ a combination of an intrinsic dimension reduction together with a graph convolutional neural network. Based on this, we plan to derive approaches for the analysis of cardiac shape trajectories to improve decision support and therapy planning.

Multiple Co-Varying parameter

A largely unaddressed topic is the extension of shape analysis to other types of data. For example, the geometrical artifacts studied within archaeology typically belong to the class of man-made objects and, thus, feature a rich geometrical, structural, and topological variability. This violates the assumption that instances of an object class can be modeled as deformation of a template shape. A promising direction for the analysis of objects with such variability is to adapt concepts from co-analysis of shape collections and topology-varying correspondence estimation.

Man-Made Objects and Beyond

One line of future work will be to extend the Riemannian mixed-effects framework to include multiple explanatory variables. This will provide the means to study shape trends that correlate to multiple co-varying parameters (e.g. archaeological changes may also correlate with excavation site and used stone material).
Since 2005, the number of annual operations in Germany has increased by approximately 25%. However, the available resource capacities of the hospitals such as operating rooms and staff remained almost equal or even decreased. In 2019 alone, there were about 17,000 open nurse positions in Germany. Some hospitals even had to shut down some of their operating rooms because they could not find sufficient numbers of qualified personnel. Other hospitals suffered from a lack of operating rooms such that they had to implement middle and late shifts to cope with the demand. Overall, today’s hospitals have to focus much more on efficient resource management to deal with potential bottlenecks. At the same time, they have to work more cost-efficiently because a large share of a hospital’s profits and expenses come from surgeries. Since all surgeries share scarce resources such as operating rooms and staff, they become increasingly interdependent on each other. These dependencies must be planned, giving them a high degree of attention. At the moment, the planning complexity can only be coped with by more human planners and a huge communication effort.

**IBOSS Project**

In order to improve the quality of the daily operating room schedules, ZIB started a BMBF-funded project “Information-Based Optimization of Surgery Schedules” (IBOSS) together with Freie Universität Berlin and the University of Paderborn in cooperation with the Charité Berlin, which is one of the world’s most prestigious hospitals. The goal of the project is to develop mathematical optimization tools that can compute daily operating room schedules of the highest quality.

There are many different indicators for the quality of an operating room schedule. The main objective for the hospital is to use the available resources as efficiently as possible. That means, in particular, minimizing overtime, idle time, delays/cancellations of surgeries, waiting time, and incorporating planning stability such that the initially planned schedule is robust enough to cope with unforeseeable events during the planning day, such as emergencies.

In particular, the different objectives are motivated by different parties such as the patient, staff, or management. Hence, it may be the case that objectives contradict each other, for example, if one has to decide whether a surgery is still performed one hour before the regular shift ends, in good knowledge that the surgery will presumably take two hours. Either the staff or the patient is going to be affected negatively. Thus, one has to weigh the different objectives to generate schedules that implement the best compromises between all the objectives.
Daily Struggle with Constraints

Around afternoon, the baseline schedule for the next planning day has to be finished. Normally, minor changes will be applied until the next morning but the baseline remains fixed. Hence, the first planning step is to create a baseline schedule. The baseline schedule already has to satisfy many side constraints. It is dictated by the OR capacity plan, which is a week-based plan that shows for how long which department is booked into which operating room. Departments describe different medical disciplines such as general surgery, gynecology, urology, or orthopedics, each of which has its own qualified employees. The OR capacity plan is valid for one year, after which adjustments are made according to current demands. The baseline schedule consists of the assignment and sequencing of surgeries to the available resources. Resources are, for example, operating rooms, surgeons, anesthetists, nurses, and medical devices.

Every surgery generates its own supply chain in the hospital system, which consists of different subphases such as station pickup, OR preparation, induction, surgery, OR cleanup, recovery, and possibly ICU. Every phase has an individual resource demand and some phases run in parallel. For example, the operating room is prepared by the nurses while in parallel the patient receives anesthesia by an anesthetist and a supporting nurse. For the surgery phase, one surgeon must be present. Often, a second surgeon will attend the surgery only during the main operation so that sometimes he/she can switch smoothly between two operating rooms.

A surgery usually requires one to two surgeons, two nurses, one anesthetist, and one anesthetist nurse. For complicated surgeries, additional staff may be required.
One goal is to minimize the delays between the subphases since every minute that exceeds the OR opening hours is considered as overtime. This is because the staff rosters are built based on the planned OR capacity. For each staff type (surgeons, anesthetists, nurses), the schedule is covered by individual shifts of different quantities, e.g. 7 a.m.–3 p.m. (6×), 10 a.m.–6 p.m. (2×), and 12–8 p.m. (2×).

Additionally, there are also individual scheduling constraints that make the construction of a plan by hand really complex and hard to grasp. For example:

- Patient p is first available at 11 a.m.
- Surgeon s is not available from 10 a.m.–12 p.m.
- Sudden absence (illness etc.) of staff
- Integration of emergencies into the running process (immediately or within 6, 12, or 24 hours) with available resources.

Sophisticated surgeries additionally require at least one very experienced surgeon so that different experience levels may affect the schedule. For each department, this requires careful planning to guarantee the best treatment for each patient. The additional planning effort comes on top of the work of the department heads, who are often surgeons or physicians themselves.

Similarly, nurses are usually qualified for one specific department. For example, it is desired that a surgery is only attended by nurses of the associated department. However, if there is a deficiency of nurses in one department, it is sometimes allowed to borrow a nurse from another department in order to finally perform the surgery. Again, this induces dependencies across the departments that are difficult to manage manually, since the departments are mainly operated separately. Furthermore, in contrast to surgeons, a nurse can be replaced by another nurse during surgery to get a mandatory break or to finish their shift on time. However, even recognizing such potential resource availabilities is a hard task because most digital real-time information is poorly formatted so that the planner cannot use it efficiently. Hence, much of the information flow still relies on direct communication.
In general, it is desired to perform the baseline schedule as planned. In practice, however, the baseline schedule will almost never be realized. This is due to the high uncertainty in the surgery durations that can sometimes vary by several hours. If a surgery is delayed or even cancelled, then replanning is necessary to keep the schedule cost-efficient. Hence, instead of considering the static cost of the baseline schedule, we want to optimize according to the dynamic cost of the baseline schedule that includes the estimated replanning decisions. Thus, we want to compute an optimal decision policy that minimizes the total expected cost with respect to the static baseline schedule as starting point.

In order to quantify the uncertainty in the system, we compute probability distributions for the duration of the subprocesses of each surgery. In general, a surgery consists of multiple smaller operations, each of which has a unique representation in form of a so-called OPS code that shows exactly which operation is being performed. For this, we developed a machine-learning model that predicts a probability distribution for surgery based on a given set of OPS codes. Our database consists of six years of historical data that contains about 140,000 surgeries. The model uses a cross-entropy objective to maximize the likelihood that the estimated distribution matches the observed durations. It turned out that most surgeries do not follow a Gaussian distribution but rather a long-tail distribution, such as a log-normal or log-logistic distribution (see Fig. 1 [3]). This seems plausible, since a surgery is more likely to take longer than usual.
Algorithmic Approach

In order to compute a baseline schedule that satisfies all the different constraints, we formulate the underlying scheduling problem as a mathematical program that is solved by a branch-and-bound algorithm based on constraint-propagation and simulation techniques. It turns out to be extremely difficult to provably compute the optimal solution. In fact, the simultaneous assignment and sequencing character of the scheduling problem is solvable only for very small instances in the proposed time frame of five minutes, even in the deterministic setting. Therefore, we combine our branch-and-bound algorithm with a primal search algorithm that heuristically cuts off unpromising branches and diversifies the search regions. The built-in constraint propagators take care of the feasibility of the solutions. Stochastic cost bounds are determined by internal simulation. Our method is able to quickly generate schedules that can be used in practice. In addition, our algorithm works online, meaning that it suggests rescheduling decisions if the current schedule turns out to be inefficient due to delays, cancellations, or no-shows. The mere time effort for the generation of a schedule is estimated to drop by at least 80% if measured in working hours. Moreover, overtime is estimated to be lowered by 17% by a better a priori evaluation of surgery durations and scheduling decisions. Our next goal is to deploy our algorithm for practical usage. However, it is a long way from an algorithm prototype to an algorithm that is used in practice by default. Data interfaces, visualizations, and UIs must be provided to run the algorithm with the correct parameters (see Fig. 2). Moreover, it is important to give the practitioners a basic understanding of what the algorithms are doing. The acceptance of the practitioners is a factor that should not be underestimated, since nobody will use even the best algorithm if they do not comprehend its benefits.

Fig. 2: Prototype of the IBOSS surgery schedule planner.
Vision of Full Autonomy

The developed algorithm makes replanning suggestions automatically so that, in the best case, the planner only has to confirm the decision made by the algorithm. This almost gives the impression of a fully autonomous system that can handle the surgery scheduling by itself - like AI. Perspectively, how realistic is the fully autonomous artificial OR planner?

Well, currently there are still many obstacles between now and that vision. First of all, current surgery planning is still far from being completely digitized. It is still based on a lot of human interaction and communication because the planning is highly decentralized. The rostering and daily scheduling for physicians, surgeons, nurses and anesthetists is mostly done by separate organizational units, but which all belong to the same supply chain of a surgery. This is historically grown but there are also few possibilities to do it differently since the manual organizational effort is already very high for all the planners. Hence, moving from decentralized to centralized planning requires the will of the hospital staff and management to be implemented practically.

A second aspect is the planning responsibility, which is a bigger hurdle for full autonomy. Similar to autonomous driving, one always needs a person that is in charge of the current events. In particular in hospitals we need to be able to act fully autonomously as a human where bottleneck decision may be the difference between life and death. Hence, it is questionable if there will ever be a fully autonomous operating room scheduling system. It is much more likely that a semiautonomous system will be implemented that will support the planner with all the necessary information and tools to run the daily operative business as efficient as possible.

(Re-)focus on the Patient

An algorithmic decision support tool yields obvious benefits to save time and money for the planners and the hospital management. However, it remains for how the patient would ultimately benefit from that change to be discussed, since all the processes should be centered around the best possible treatment for the patient. We believe the benefits for the patients are manifold.

An algorithmic tool is capable of making objective decisions based on the given real-time data. There is probably nothing more annoying for a patient than traveling to the hospital for surgery, waiting for several hours, and finally getting a notification that the surgery has been cancelled. Prescriptive analytical tools have all the information available: the status in the operating rooms, the expected waiting time of each patient, and the likelihood of cancellation, even hours before the actual appointment. If a critical value is exceeded, the patient may get an earlier notification of cancellation or the surgery can still be performed by an improved resource management.

Furthermore, a smaller number of planners lowers the danger of miscommunication and potential conflicts between staff in the operating rooms. The less personal communication is necessary, the more the staff can focus on their work. Current polls show that about 50% of the OR staff feels more stressed because of personal conflicts at their workplace, which they consider as a serious risk for the patient. Less necessary communication avoids unnecessary personal conflicts. This has immediate consequences on the treatment quality of the patient, since the whole working atmosphere in the operating room can be improved by using a computational support tool in the background. This will probably lead to best improvement in the treatment quality of a patient, since the most important decisions in the operating room are still made by human beings.
Unbeknownst to most, mathematical optimization is ubiquitous in our daily lives. Used in most large companies and organizations in the world since the advent of Industry 3.0, it optimizes manufacturing and services drastically, and will become even more prevalent through Industry 4.0. Cutting-edge mathematical research and software are critically needed to tackle these new challenges. ZIB is at the forefront of these efforts, leading research and continuously improving the in-house solver SCIP, one of the fastest academic mathematical optimization software packages.
Determining the best course of action in a given situation is a fundamental task performed by humans on a daily basis. In the modern world, it often involves complex, large-scale decisions: What is the best use of limited production resources? How to construct a schedule that achieves maximal efficiency while avoiding conflicts? Which connections should be added when expanding a power network so that the costs and environmental impact are minimized? These decisions may involve thousands and even millions of variables, making computational tools necessary in order to obtain high-quality solutions. Providing such tools is the aim of mathematical optimization, which is the science of finding the best choice under given restrictions.

A mathematical optimization problem is described in terms of variables, which represent decisions and their consequences, and parameters, which reflect the knowledge about the system to be optimized. The goal is to find the maximum or minimum of an objective function (in some cases, several objective functions) while satisfying constraints. Some common types of constraints are equations, inequalities, logical relations, variable domains, and requirements that certain variables have only integer values.

Graph of a nonlinear function. The global maximum is achieved at point A (image created in GeoGebra).
One well-known optimization problem is the knapsack problem, which has numerous applications such as portfolio optimization, warehouse space management, and raw material cutting, to name a few. The name is derived from an everyday situation where a traveler must decide which items to pack into his or her knapsack by considering their usefulness and weight.

More formally, a set of indivisible items is given, where each item has a given weight and value. The combined weight of the chosen items is constrained to be below a given weight limit. Each item has a variable associated to it, encoding the decision of whether to choose this item (variable has value 1 in the solution) or not (value 0). No fractional solutions are allowed, since the items are indivisible. The objective is to maximize the combined value of the chosen items. A slightly different version of this problem considers types of items, and the variables represent the numbers of items chosen from each type.

This is a combinatorial optimization problem because it deals with a finite set of objects. The main difficulty here lies in the fact that the number of possible combinations that comprise a feasible solution grows exponentially in the size of the problem.

Combinatorial optimization is one of several interesting and challenging areas of optimization. Other common problem classes include mixed-integer linear programs (MIPs) that contain both discrete and continuous variables, nonlinear programs (NLPs) where some constraints are described with the use of nonlinear functions, and the combination of these two classes, mixed-integer nonlinear programs (MINLPs).
Given the practical relevance of optimization problems, researchers at ZIB have developed the general-purpose solver SCIP to solve large-scale problems to global optimality. SCIP is a framework for constraint integer programming oriented towards the needs of mathematical programming experts who want to have total control of the solution process and access detailed information down to the guts of the solver. SCIP can not only be used to solve difficult optimization problems, but also serves the optimization community to implement, test, and evaluate their algorithmic ideas in a state-of-the-art solver. Indeed, with over 14,000 downloads per year from over 100 countries, SCIP is one of the most used research frameworks for branch-cut-and-price algorithms worldwide – also because SCIP is currently one of the fastest MIP and MINLP solvers that is openly accessible in source code. Additionally, starting from version 3.2, SCIP’s main new developments and features are documented in release reports [1, 2, 3, 4].
A visualization of SCIP’s plug-in-based design.
Interfaces

A natural way of formulating an optimization problem is to use a modeling language. Besides the modeling language ZIMPL, which is developed at ZIB, there are several other modeling tools with a direct interface to SCIP. These include Comet, a modeling language for constraint programming, AMPL and GAMS, which are well-suited for modeling mixed-integer linear and nonlinear optimization problems, and CMPL for mixed-integer linear problems. Additionally, SCIP provides interfaces to several programming languages that are developed and maintained on GitHub in order to provide extensions and patches faster and to collaborate on them more easily. Besides the popular interfaces for Python and Java, there is also an interface for Julia available.

Development and Quality Control

With more than 500,000 lines of code, the source code of SCIP is not trivially maintainable. To facilitate and organize the development, Git and GitLab are used as a source code management system. The code is organized in branches on which bugs are fixed or features are developed that are then merged to the main master and bug-fix branches. Git decentrally and hierarchically preserves the history of code changes. GitLab provides a platform for discussion, including description of bugs and issues. Ensuring code quality is of the highest importance; for this we use a Jenkins server that regularly runs our test suite on different platforms and environments. Additionally, before applying a code change to one of the main branches, a short series of checks is automatically executed and results are reported back to the developer. Finally, the continuous improvement of the performance of SCIP is ensured by regular performance runs. These consist of the execution of SCIP on the compute cluster on a set of problem instances. The log files are scanned for different measures, for example speed and solution quality. Since the evaluation of this data can be a difficult task, a custom software tool, Rubberband/IPET, is used to store the log files, collect the data, and facilitate and standardize the analysis. The components interact with the developers as well as with each other on demand and combine to form an elaborate system for quality control.
CAREERS

- Ten awards for master’s and Ph.D. theses at MOS, EURO, GOR, and DMV
- Nine former developers are now building commercial optimization software at CPLEX, FICO Xpress, Gurobi, MOSEK, and GAMS

RECENT AND PLANNED SCIP WORKSHOPS

UG workshop
Workshop on parallel algorithms in tree search and mathematical optimization:
- 15 speakers
- 46 registered attendees
- 16 talks
- Two programming sessions

Google × SCIP workshop
Purpose: For Google engineers to better understand how SCIP works, for SCIP developers to better understand how Google uses MIP, and for both parties to recognize each other’s priorities and identify opportunities for future collaboration.
- 11 lectures with discussions

Combinatorial optimization at work 2020
- Two-week summer school to be held online on September 14–26, 2020
- 27 confirmed speakers at the moment

Online SCIP workshop 2020
To be hosted by the University of Exeter on June 3–4, 2020
SCIP is the main component of a toolbox for generating and solving mixed-integer nonlinear programs. The SCIP Optimization Suite consists of SCIP, SoPlex, ZIMPL, UG, and GCG. The SCIP Optimization Suite is the result of a continuous collaboration between researchers located at ZIB within the context of SynLab, one of the labs of the Research Campus MODAL, at research centers across Germany, such as TU Darmstadt, FAU Erlangen-Nürnberg, RWTH Aachen, as well as other countries in Europe such as the UK and the Netherlands. Overall, the development team consists of 11 postdocs and professors, 13 Ph.D. students, and five student assistants.

The SCIP Optimization Suite

The Power of Collaboration

SUCCESS STORIES AND APPLICATIONS

Large-scale mathematical optimization problems arise in various projects developed at ZIB. Here, we present a selection of projects for which SCIP was critical for their success.

ZIMPL
Zuse Institute Mathematical Programming Language (ZIMPL) is an algebraic modeling language. ZIMPL supports the creation of linear as well as nonlinear mixed-integer mathematical programs.
https://zimpl.zib.de/

SoPlex
Sequential object-oriented simpLex (SoPlex) is an efficient simplex-based linear programming solver. SoPlex allows linear programs to be solved exactly.
https://soplex.zib.de/

GCG
Generic Column Generation (GCG) is a generic branch-cut-and-price solver for mixed-integer linear programs. Based on SCIP, GCG solves MILPs using Dantzig-Wolfe decomposition. The decomposition is based on a structure provided by the user or detected by GCG itself.
https://gcg.or.rwth-aachen.de/

UG
Ubiquity Generator (UG) is a framework for parallelization of branch-and-bound solvers in a distributed or shared memory computing environment. UG has been used to create massive parallel solvers by parallelizing SCIP, CPLEX, and Xpress. ParaSCIP solved 21 open instances from different MIPLIBs. ParaXpress solved three open instances from MIPLIB2010 and one from MIPLIB2017.
https://ug.zib.de/
The SAP Collaboration

Supply chain management (SCM) deals with the optimization of a company’s value chain. This begins with the choice of suppliers and extends to the delivery of produced goods to the customer. In this process, decisions are made about purchasing, transport, storage, and further processing of the goods. The value chain is depicted symbolically in the figure below.

A visualization of a supply chain management instance.

Production planning plays a central role here: based on customer orders or the forecasted demand, it decides when and where which products are produced. In doing so, it ensures that the materials required for production are available in sufficient quantities at the respective location, takes into account capacity restrictions, such as the number of available workers, and attempts to reduce storage costs by producing as late as possible.

The integrated planning of the entire value creation process contains a high potential for optimization and is therefore of great importance for the profitability of companies. Our industry partner SAP is developing software that enables companies to model their value chain in detail. This includes precise descriptions of production models, capacity restrictions, as well as costs of production, storage, transport, and delivery. In addition, fixed incoming stock, customer orders, and demand forecasts are specified for different time periods (for example, days or weeks).

A mixed-integer linear optimization problem (MIP) is then generated from this data to optimize the supply chain.

The SAP project deals with the fast solution of especially numerically demanding and large MIP instances that arise from such supply chain management problems. Thereby the primary goal is not the optimal solution, but rather the best possible solution within fixed time limits. The SCIP Optimization Suite, which was extended with new components added during the course of the project, is used to solve the MIP instances.

The complexity and the large size of the problems motivated several SCIP methodological improvements [5]. For example, several problems exhibit underlying structures that could allow for more efficient, decomposition-based solution algorithms. A special decomposition API is thus being developed within SCIP in order to read user decompositions in a more flexible way and allow for better structure exploitation in the developed algorithms.
### Energy Grid Berlin-Adlershof

As part of the energy strategy Berlin-Adlershof 2020, the high-tech location Adlershof has set itself the target of reducing its primary energy demand by 30 percent by 2020. As part of various projects, several energy-related concepts and measures have been implemented at the Adlershof campus. The project “Energienetz Berlin Adlershof,” implemented as a joint project of the three partners TU Berlin (TUB), Siemens AG (in collaboration with ZIB), and Hochschule für Technik und Wirtschaft Berlin (HTW Berlin), made a significant contribution to achieving this goal. The project aimed at improving energy efficiency at the property and district level as well as creating planning bases for the efficient energy supply of urban neighborhoods. The project demonstrated a networked energy system that included heat, cold, and electricity as forms of energy. Based on this supply system, a cross-media energy management system has been implemented.

In this context, optimization problems need to be solved, which can be formulated mathematically as mixed-integer programs with quadratic constraints, MIQCPs for short. Quadratic constraints become necessary when accurately modeling nonlinear physical processes, in this case, the heat transfer between ice storage, heat storage, and the connecting thermal networks. To be able to treat the projected detail levels in realistic problem dimensions, the performance of current MIQCP algorithms had to be improved. To this end, the MINIP solver SCIP, developed at ZIB, has been further developed to generically exploit emerging structures such as bilinear mixing conditions and time-indexed decision variables. Performance on the given MIQCPs has been greatly improved, and SCIP was successfully integrated into the energy management system, helping to plan and improve the given energy system.

### plan4res

The general objective of the EU Horizon 2020 project plan4res is to fill the gaps between the increasing complexity of the future energy system planning and operational problems, and the currently available system analysis tools. In a cooperation of seven partners from research and industry, enhanced end-to-end planning and operational tools dealing with technological and market uncertainty, emerging technologies, and increased sector coupling of multienergy vectors, such as heat, cold, and transport, will be assembled. This would result into a synergistic approach to support European system planners, operators, decision-makers, and regulators.

plan4res [6] requires a holistic approach to accomplish its ambitious objective. This approach results in very challenging integrated models that have a high level of detail at European scale with precise spatial and temporal resolutions and that take uncertainty into account. Solving these models requires using decomposition and aggregation methods together with improved solvers. The SCIP Optimization Suite is going to be used by plan4res’ planning and operational tools to solve the LPs and MIPs arising from optimization models of energy systems, either separately or as part of a large stochastic optimization methodology.

---

![Schematic topology of a smart grid with various energy mediums.](image1)

![Ice storage at the Adlershof site.](image2)

![The logo of the plan4res project.](image3)
The GasLab: Gas-Flow Prognosis

In the framework of the Research Campus MODAL, the GasLab addresses the main challenges of natural-gas dispatching in light of the German energy transition (Energiewende) and the crucial role that Germany plays as a transit country in the European cross-border natural-gas supply. The research is conducted in a close partnership with Open Grid Europe (OGE), one of Germany’s largest transmission system operators, who is responsible for operating about 12,000 kilometers of natural-gas pipelines in Germany: roughly the same length as the overall German Autobahn network. Through large-scale software, real-time decision support is provided incorporating mathematical models that depict the technical aspects as well as the nontechnical operational measures of the gas transmission network. Additionally, machine-learning techniques are thoroughly investigated, alongside efficient solution algorithms for the respective optimization problems.

An important component in the GasLab is to improve OGE’s hourly forecasts for gas demand and supply. The goal is twofold: to predict as precisely as possible the average hourly gas flows for the upcoming gas day and to overcome the diversity in the data characteristics in order to deliver predictions that are appropriate for all different types of entries and exits (e.g. connections to other networks/countries, industrial users, municipal consumers, etc.). A multiregression prognosis methodology is developed for this purpose by training a prediction model using a comparably short historical time. In order to accurately describe the given historical time series, a function with predefined feature terms is formulated by optimizing over its coefficients (weights) to reach the smallest error over the training horizon. Moreover, in order to avoid the potential risk of overfitting, that is computing a function that perfectly represents the past but is very poor in predicting the future, the prognosis method is further extended to include coefficients’ selection decisions. This development ensures that at most k coefficients of the prediction function are nonzero, by selecting the most significant ones. Indeed, this higher level of precision comes at the expense of introducing more functional terms and discrete decision variables to the optimization model; thus, solving a large-scale mixed-integer linear program.

SCIP is currently used at the base solver for this vital optimization problem. The models are solved to optimality within an acceptable time frame for the industrial requirements. Furthermore, in case of symmetric solutions, it has been noted that SCIP delivered the solutions that perform best in terms of prediction accuracy for this application. This optimization-based multiple regression method can now be used for predicting the future values of all kinds of time series.
Many optimization problems arising in practice are based on an underlying (physical) network. Consider, for example, the construction of telecommunications networks. Operators are faced with the task of connecting customers and deciding where to route new cables in the most cost-efficient way. Similar scenarios are also found in many other applications. They lead to one of the most studied problems in combinatorial optimization and computer science, the Steiner tree problem in graphs. Mathematically, such problems can be modeled by graphs, abstract representations of “networks.” A graph consists of a set of so-called vertices (corresponding to locations) that are connected or can be connected by so-called edges (cables). One can now associate a non-negative weight (distance or cost) with each edge. A subset of the vertices is designated as terminals (customers). A tree – a connected subgraph without any cycles – is called a Steiner tree if it connects all terminals. One is usually interested in computing a Steiner tree of minimum length or cost. This seemingly simple problem called the Steiner tree problem in graphs, or SPG for short, has proven to be notoriously hard to solve, both in theory and in practice. Indeed, simply checking each possible Steiner tree and choosing the best among them becomes quickly intractable; for problems with a few hundred edges, this approach would already take millions of years on the fastest currently available computers.

Moreover, many variations and generalizations of the SPG arise from practical applications. The diverse fields where one encounters such problems encompass computational biology, computer chip design, computer vision, the deployment of drones, and recently also machine learning. To efficiently solve many of these problems, the Steiner tree framework SCIP-Jack has been developed at ZIB as part of the SCIP Optimization Suite. SCIP-Jack can currently handle the classical SPG and 13 related problems. Despite this versatility, SCIP-Jack is the fastest solver for most of the 14 problems it can solve. For example, at the PACE Challenge 2018, dedicated to so-called fixed-parameter tractable SPGs (a subclass of SPGs), SCIP-Jack was the winning solver in one category, and took second and third place in the remaining two. Even though SCIP-Jack (unlike the other participants) does not provide any specialized routines for these subproblems. Steiner tree problems can also be modeled and solved as general MIPs. However, the highly specialized algorithms used in SCIP-Jack perform much better: problems with a few thousand edges can usually be solved within seconds by SCIP-Jack, but take weeks to be solved even by the fastest commercial MIP solvers – or cannot be solved at all. The largest problem solved by SCIP-Jack so far, derived from a cancer application, has more than 64 million edges.

A real-world telecommunications network from an Austrian city with an optimal Steiner tree as a solution marked in red.
Exact IP/LP

The vast majority of linear and mixed-integer solvers rely on floating-point arithmetic, due to its superior computational capabilities. The occurring small imprecision is controlled by the careful handling of error tolerances. In most real-world applications, it is sufficient to find a result that is no more than one millionth from the exact optimum. However, there exist problems with the requirement of an exact optimal solution, without any numerical imprecision. The exact versions of SCIP and SoPlex achieve this by employing a combination of fast floating-point arithmetic and slower but exact symbolic computations.

Use cases that require exact solutions stretch from research to industry applications that do not allow for any violations of the problem constraints. Exact linear programming was used in molecular biology where multiscale reactions led to problems that could not be handled by floating-point solvers [7]. This sparked the development of the exact solving mode in ZIB’s linear solver SoPlex. Exact SCIP was developed and used at ZIB to investigate two famous open conjectures formulated by Péter Frankl [8] and Václav Chvátal [9], respectively. On the industry side, in the design verification of integrated circuit designs, exact solvers are necessary to avoid the accumulation of errors in the numerous different components [10].
MACHINE LEARNING AND BIG DATA
Big data (BD) and machine learning (ML) (together with decision-making) are the key pillars of artificial intelligence (AI). The analysis of very large and heterogeneous amounts of data has the potential to revolutionize many areas of our lives, from the sciences, production, transport, and energy, to political and social processes. However, dealing with BD and ML requires highly specialized knowledge and infrastructure and poses many challenges: it requires computer skills, mathematical skills, and engineering skills. It further demands a societal discussion regarding its ethical implications. BD and ML will disrupt all levels of society and will inspire completely new applications fueling innovation and strengthening the economy. The impact is already evident today: we entered the “fourth paradigm” in the sciences and the economy is talking about the next industrial revolution.

Research at ZIB

Research in ML and BD has been an integral part of research activities at ZIB for ten years now. With recent success stories and the strong growth of international research activities, ML and BD are becoming a cornerstone of our research strategy. Therefore, in fall 2019, we decided to create a new research department, “AI in Society, Science, and Technology,” at ZIB that is strongly integrated with our other departments. In addition, a new Berlin-based Competence Center, the “Berlin Institute for the Foundations of Learning and Data” (BIFOLD), was founded that integrates the former Berlin Center for Machine Learning (BZML) and the Berlin Big Data Center (BBDC) into a new structure that is funded by the German Federal Ministry of Education and Research (BMBF). ZIB, having been part of both, BZML and BBDC, is part of the consortium supporting BIFOLD and will implement new research groups in this context. In the following, we report on four sample research projects on BD and ML that show the breadth of our research activities in these fields and which we are going to strengthen and intensify.
Decision Support for Gas Network Operation

Would you feel comfortable driving a car using your rear-view mirror and your driving experience only? Open Grid Europe (OGE) is a transmission system operator responsible for the delivery of more than 25% of the German primary energy consumption by operating a natural gas network of comparable length to the German Autobahn. To meet all demands, the 24/7 dispatching center operates more than 100 compressor units, almost 300 regulators, and more than 3,000 valves to control the network. Still, the network is operated mainly based on measured data, which is only available in the "rearview mirror," and the expert knowledge of the dispatcher. To improve this situation and to anticipate and prevent critical situations in the network, researchers at ZIB, together with a group of experts at OGE, have developed a smart, forward-looking, analytics-based decision support system. For this to work, it was necessary to utilize three types of analytics:

- Descriptive: modeling and simulating the gas flow in the network
- Predictive: predicting future gas supply and demand at the entries and exits of the network
- Prescriptive: recommending network control measures to ensure safe and efficient operation of the network

The GasLab in MODAL

In the early 2000s, the European Union changed the regulative framework for the gas market toward fulfilling the main European energy objectives: competitiveness, security of supply, and sustainability. In the unbundled gas market, demands, supply, and storage facilities are beyond the control of the transmission system operator. As part of the decarbonization of the energy system, short-term transport demands for the rapid supply of gas-fired power plants as a supplement to fluctuating electricity generation from renewable energies are increasing sharply. These fluctuations require the dispatching center to have quick reactions, which can only succeed if the network is prepared to deal with such short-notice transports. To tackle this challenge, five years ago, together with OGE, ZIB started the GasLab as part of the Research Campus MODAL, funded by the German Federal Ministry of Education and Research. The goal of the GasLab is to develop a decision support system that predicts the future gas flows, analyses control options and risks, and intelligently computes recommended actions to secure safe operation while minimizing energy consumption and equipment wear.
Machine Learning Joins Optimization for Decision Support

Every 15 minutes, the optimization core computes recommendations based on the current state of the network, its past, and its technical capabilities. A single station in the network can have more than 1,250 discrete operational modes. Each mode may include target values for continuous quantities such as the target pressure of a regulator. First, an hourly forecast for the more than 1,000 entry and exit points of the network for the next 24 hours is computed employing a mixture of machine learning and optimization on a preselected set of features most suitable for each entry or exit. Based on these forecasts, the recommended operational measures are computed. However, an exact transient model, including all discrete and continuous variables for the full network is computationally intractable. Therefore, we employ a two-phase approach, decomposing the different aspects of the problem by exploiting the topology of the gas network: a coarse model computes amounts and directions of flows, computing when to transport how much gas on which path through the network. Then, detailed models for the individual compressor stations are solved in parallel. These validate and complement the amounts and directions calculated by the coarse model and compute precise action recommendations. The objective is to fulfill demands while minimizing costly mode changes considering many additional constraints needed to ensure the practical feasibility of the recommended action. Finally, the results are summarized and displayed to the dispatchers on dedicated iPads, providing them with a set of directions.
Consistent State Management for Data Stream Analysis

Big Data Stream Processing

Data stream processing frameworks such as Flink [1], Spark [2], and Hadoop require the data analysis pipeline to be expressed as a directed dataflow graph. Unfortunately, this is sometimes difficult or even impossible. In some cases, the flow graph is not static but changes its pathways depending on the input data. Despite much research in science and industry, data access and the data handling in stream processing systems is still based on a few, overly simplistic interfaces such as parallel file systems, the Hadoop Distributed File System HDFS, or the Google File System GFS [3, 4, 5].

None of these systems allows to communicate nonfunctional properties of the pending data accesses. Relevant nonfunctional properties include data access patterns, access concurrency, repeated data accesses, required data redundancy, location restrictions, or similar aspects. Clearly, data stream processing could be made much more flexible, portable, adaptable, and robust against single component failures if the nonfunctional properties were expressed before starting an analysis job and taken into consideration by the stream processing environment at runtime.

Fig. 5: Declarative API mapping requirements to executable implementations, with three mappings in red, blue and green.
Consistent State Management – What Is It Good for?

For the current and future big data analysis pipelines with further increasing complexity, it becomes more and more clear that the concept of independently treated data streams leads to workflows with unnecessary synchronization and inefficient resource utilization. With a well-designed and properly implemented consistent state management, stream processing can be much enhanced, leading to better utilized resources and faster job turnaround times. All subtasks of an analysis workflow must be allowed to access the shared state, or a part of it, in order to exchange information among themselves, such as threshold values for search algorithms, previously learned classifiers, or similar.

The consistent and efficient management of the concurrently accessed shared state information, which is distributed over all accessed resources, is a difficult problem – both in theory and practice. Especially in the context of multiple, potentially widely distributed and heterogeneous resources and the calculation on-site, new challenges arise.

In high-security applications like information-based medicine, for example, the data must be processed and analyzed on local resources at the respective institution (clinic or laboratory) and may only be further processed as aggregated, anonymized data. Even so, it should be possible to derive meaningful analysis results by combining aggregated data across the various resources of the participating institutions with better representativeness of the total number of cases and to avoid or consider a possible bias of the sources of an institution.

Research Questions

Consistent management of distributed state information is at the core of distributed algorithms. Several requirements and research questions arise from this problem:

1. Where should data be placed in order to enable the efficient processing of iterative algorithms in geographically distributed environments?
2. Which description of data locality, degree of aggregation, and anonymization is required?
3. How can failure tolerance and distribution restrictions be specified and realized?
4. How can heterogeneity, dynamic data-space partitioning, self-management, scalability, bandwidth utilization, and incremental data modification be efficiently integrated and implemented?

To answer these questions, we draw on our research and development experience with the two distributed data management systems developed at ZIB: XtreemFS, a failure-tolerant distributed file system, and Scalaris, a distributed failure-tolerant NoSQL database. A suitable declarative abstraction layer will be developed for various requirements of state management, e.g. regarding failure tolerance, data placement, I/O characteristics, and scalability, which will then be realized automatically with the available resources and transparently optimized for the application.
Our improved Paxos algorithm with in-place consensus sequences [6] will serve as a basis for new algorithms. Additionally, modern hardware features such as one-sided communication with remote direct memory access (RDMA) and persistent memory (nonvolatile memory, NVM) will play an important role in deriving new solutions. While all these new hardware features may be beneficial for specific computer science systems, they pose additional challenges for distributed systems, such as atomicity and visibility of operations with concurrent accesses in the face of compute nodes operating under the crash recovery failure model.

Extracting Dynamical Laws from Data

In all areas of science, technology, and society, vast amounts of data are becoming available that can be utilized to analyze, control, and understand the underlying complex processes. The past decade has seen tremendous progress regarding algorithmic methods for data-driven prediction of processes. In classical machine-learning settings, it is often acceptable to have a black-box prediction method, provided that it performs efficiently and with small error. In science and technology, the essential objective is often to obtain an interpretable model that leads to a fundamental understanding of the observed process. One aims to derive a theory that allows for successful predictions in regimes that are very different from those in which the underlying data was gathered. Estimation methods that can generate understanding and deliver physically interpretable and mathematically tractable models from data are still in their infancy.

The real challenge lies in inferring effective equations in mesoscopic regimes of complex systems, where the microscopic dynamical laws are not known or do not lead to a practical description and human intuition is often overwhelmed by the sheer complexity of the system. Examples include developmental biology, where the development and cell differentiation of an organism can be tracked by time-dependent microscopy with single-cell resolution, or social or societal processes, where, in most cases, a fundamental theory that would allow a model for the dynamics to be derived, are simply not available. At ZIB, our aim is to combine machine learning and mathematical process simulation in order to derive effective dynamical laws and to obtain models that generate understanding and insight.
Data-Driven Transfer Operator Techniques

Data-driven approaches for the analysis of complex dynamical systems – be it methods to approximate transfer operators for computing metastable or coherent sets, methods to learn physical laws, or methods for optimization and control – have been steadily gaining popularity over the last years. Algorithms such as EDMD [7, 8], SINDy [9], and their various kernel-, tensor-, or neural-network-based extensions and generalizations have been successfully applied to a plethora of different problems, including molecular and fluid dynamics, meteorology, as well as mechanical and electrical engineering. Most of the aforementioned techniques turn out to be strongly related, with the unifying concept being transfer operator theory, whether Koopman or Perron-Frobenius or their stochastic variants.

Inferring dynamical laws from data mostly means estimating the right-hand side of a (mostly stochastic, perhaps partial) differential equation given rich enough time series data. SINDy constitutes a milestone for such a purely data-driven discovery of dynamical systems. Its main idea is to approximate the right-hand side of the differential equation via a sparse linear combination of terms coming from a rich library of ansatz functions. While SINDy performs very reliably for deterministic dynamics, there are some pitfalls for stochastic settings or large amounts of noise. In order to overcome these obstacles, we developed a method for estimating the generator of the Koopman operator from data [10], based on an ansatz space spanned by a rich library of nonlinear functions as in SINDy.

First Results

To this end, in one of our BIFOLD projects, we developed a general framework for computing an approximation of the Koopman generator, both for deterministic and stochastic systems, and started exploring a range of applications [10]:

We illustrated that the governing equations of deterministic as well as stochastic dynamical systems can be obtained from empirical estimates of the generator.

In addition, we explored two powerful applications of the approximated Koopman generator. We showed that the resulting method, termed gEDMD, can be used to identify coarse-grained models based on data of the full system, which is a highly relevant topic across different research fields, like molecular dynamics simulations for instance. Moreover, we apply the Koopman generator to control dynamical systems, providing flexible and efficient model predictive control strategies.

In the next step, we will investigate how gEDMD can be applied to social and societal processes, e.g. in opinion formation or mobility decision processes.
Stream Processing with Blockchain Technologies

Auditability and Reproducibility in Stream Processing

Reproducibility and traceability play a central role in justifying and supporting credibility not only in science but also in business applications. To realize both for data-intensive research applications, for example in the field of machine learning, that handle dynamic unbounded data streams, several methodological and technical challenges have to be solved.

Traceability for small data sets can be reached with the promising technology of blockchains, which store and distribute intermediate states of a sequence of data manipulations so that they cannot be tampered with by individuals, as in a ledger. To process large amounts of data without the digital ledger growing to an unmanageable size, data can be stored off chain. The challenge here is to protect the off-chain data from manipulation. It becomes even more difficult if not only the processing of static, closed data objects is to be made credible and auditable, but also the results of a continuous data analysis on data streams, which is to be made credible and auditable up to a given point in time. For this problem, no suitable solutions are known yet.

Ongoing Research

We are developing a distributed data management system that supports and enables traceability and reproducibility of the results of distributed continuous data analysis on data streams. To achieve this goal, we are coupling blockchain technology with cluster data management. More specifically, we are researching, developing, and evaluating methods for manipulation protection of off-chain data in distributed environments, including algorithms for failure-tolerant, coordinated snapshot generation for distributed data stream processing. Detection and reliable deletion of data that is no longer required is another difficult task that relies on distributed snapshot processing. Our work gives answers to the following research questions:

1. How can continuously running distributed data stream applications be made auditable and reproducible?
2. How can blockchain technology and distributed cluster data management be combined to enable tamper-free, consistent snapshots on a group of data objects (e.g. files), to support distributed data stream processing applications? The application should be influenced as little as possible, i.e. the coordination of a corresponding snapshot creation should be asynchronous.
3. How and when can data objects that are no longer referenced be cleaned up automatically in order to keep storage requirements at a moderate size? Time specifications for deletion periods could also be taken into account.
4. Which data structures and layouts are suitable to enable efficient access to the latest state as well as to snapshots with low storage overhead in respect to storage capacity?
WHY IS THERE A NEW WAVE OF ATTENTION FOR FPGAs?
Currently, reconfigurable computing with field-programmable gate arrays (FPGA) is experiencing a new wave of attention, particularly in the HPC community. On FPGAs, highly data-parallel architectures can be implemented to target more power-efficient solutions. Here are some of the challenges of “programming” an FPGA.
Reconfigurable computing is a class of custom computing where algorithms are not implemented in software for a fixed logic device like a CPU but on a device, which allows high-level logic functions to be defined by the customer depending on the algorithmic needs. Field-programmable gate arrays (FPGA) are such devices enabling a custom circuit design that can be loaded within microseconds – the configuration step. This ability to reconfigure an FPGA enables the power-efficient implementation, since only the functions are configured that the algorithm really needs.

Since the birth of FPGAs in the mid-1980s, their capabilities have grown exponentially and more features have been added over time. FPGAs made their way into communication and encryption systems. Not surprisingly, within the last decade, FPGAs have become attractive for system providers and software developers seeking for energy-efficient operation and solutions, respectively. Life science (in particular genomics), search engines, high-frequency trading, or further user-defined functions near communication paths become new application domains beyond the traditional FPGA success stories. Wherever a low latency and limited power are critical demands, FPGA might be the right choice as a target platform. IoT applications (edge computing) and self-driving cars are recent examples. FPGAs are shining with highly data-parallel algorithms, since the lower clock rate of some hundreds of megahertz needs to be compensated.

Modern FPGAs provide basic logical elements like lookup tables (LUT), flip-flops (FF), and block RAM (BRAM) but increasingly also hard-code functional units like DSPs or CPU cores. Furthermore, interfaces to the chip-surrounding infrastructure as memory (DRAM, SRAM), communication (Ethernet, PCIe), and XXX require corresponding IP core that are integrated by the design tools on demand onto the FPGA. This way, in particular large FPGAs converge to system-on-a-chip (SoC) devices. Fig. 1 gives an overall schema and Table 1 summarizes key features of a modern FPGA suitable for high-performance computing and data analytics.
Typical FPGA platform infrastructure for HPC and DA workloads include multiple memory banks and a host interface, e.g. PCIe for system integration. A unique feature of FPGA devices is multiple serial links for low-latency, high-bandwidth communication, which is attractive for custom configurable high-speed interconnections linking multiple FPGA platforms or providing communication links to external data sources (experiments, IoT).

For selected workloads and steps in computational workflows reconfigurable computing can help to implement energy-efficient solutions. This way, HPC configurations with FPGA partitions can help in exascale systems to offer the required computational power while reducing the energy demands. To extend the usage profile of FPGAs to a broader range, i.e. making reconfigurable computing accessible to the majority of HPC software developers, one has to tackle the challenge of “programming” this device class. Some of the approaches are described in the next sections.

<table>
<thead>
<tr>
<th>Key Features of a Modern FPGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>900,000 configurable logic blocks</td>
</tr>
<tr>
<td>5,760 DSPs for fixed point and IEEE 754 single precision floating-point operations</td>
</tr>
<tr>
<td>11,721 SRAM blocks, configurable in data width and number of access ports</td>
</tr>
<tr>
<td>Up to 96 serial transceivers with up to 28.3 Gbit/s</td>
</tr>
<tr>
<td>Integrated DDR4 memory controllers</td>
</tr>
<tr>
<td>Typical design frequencies 300–600 MHz</td>
</tr>
<tr>
<td>Typical power consumption 50–225 W</td>
</tr>
</tbody>
</table>

Table 1: Typical key features of a modern FPGA (here an Intel Stratix 10 GX2800 FPGA).
Programming FPGAs – Really?

With so many interesting features, it is a surprise why FPGAs are not in widespread use. In contrast to CPUs, where, over the last six decades, a hardware abstraction via high-level programming languages and powerful compilers evolved, migrating an algorithm onto an FPGA means circuit design (a special engineering domain). Typical hardware description languages (VHDL, Verilog) have to be implemented and are still mainly used by engineers who want to squeeze the last bit of performance out of the hardware. But, in the last decade, approaches have become available that provide a higher abstraction level to the underlying hardware like high-level-synthesis (HLS) languages or recently OpenCL.

HLS languages are typically C/C++ derivatives supporting a subset of the C/C++ language-plus-vendor-specific compiler directives to help the C-to-HDL compiler in the mapping and optimization process. More details about the inner workings of an HLS is presented in the next section.

OpenCL [1] is a framework that provides a standardized API and runtime system for heterogeneous platforms. Together with the vendor-specific and platform-dependent infrastructure, it provides one of the highest abstracted programming models and a convenient approach to program FPGAs. With OpenCL, the software approach to "programming" an FPGA is today becoming the dominant method for a wide group of developers. Fig. 2 schematically classifies performance versus development time for the different programming approaches.

---

**Fig. 2:** High-level synthesis (HLS) aims to reduce development time compared to traditional RTL (register-transfer level) circuit design approach. Using the OpenCL approach, it costs less time to synthesize the first working design to FPGA, but more time-consuming test, debug, and optimization cycles are needed before the performance is comparable to an HLS design (idea taken from [2]).

---

One could imagine that bringing an algorithm onto an FPGA is now, with these high-level approaches mentioned above, comparable to the common software development cycle for CPUs and GPUs. Unfortunately, there is still a very time-consuming step that is part of all toolchains present in the chosen programming language: the so-called place-and-route (P&R) step. The P&R step can take hours, even when using multiple processing threads on modern hardware. Fig. 3 shows an overall view of the traditional and the OpenCL design and tool flow.

The place-and-route step is about placing the higher functional logic blocks on the FPGA and connecting these blocks in the most achievable, optimal way so that the design requirements such as highest possible frequency (shortest/longest routing path) or smallest possible area usage can be met.

To shorten the development process for program developers (and circuit designers), software emulators running on standard hardware for incremental code development and functional debugging are used. Only at certain stages of the last time-consuming step, the generation of the FPGA configuration file (bitstream) is triggered.
High-Level Synthesis – a Look Under the Hood

In order to use FPGAs as accelerators for HPC, it is important that they can be programmed by software engineers without extensive hardware design knowledge. This is where high-level synthesis (HLS) comes into play, the translation of a high-level language (HLL) like C/C++ into a hardware description language (HDL), which can be further synthesized to configure the FPGA accelerator.

Even though the compiler is capable of many automatic optimizations, as for every other target (e.g. CPU or GPU), it is important to design and express the algorithm in a manner that allows the advantages of that platform to be utilized. To be able to accomplish this and to gain trust in the compiler, one needs to roughly understand the inner workings of the HLS.

Most modern HLS tools (e.g. Intel and Xilinx) are LLVM-based, they will use a pre-existing compiler front end (e.g. Clang for C/C++) to translate the HLL into a target-independent intermediate representation (IR). The LLVM IR is similar to a typed RISC assembly and it follows the static single assignment (SSA) form. This form will assign each variable only once and create a copy if it is further modified. It is meant to simplify dependencies analysis but also suits the FPGA as a target very well, as there is no strictly limited set of registers (like with traditional CPUs) and operations are often just chained without saving them to registers in-between.

On the IR-level, a set of pre-existing and custom optimization paths will be executed that fit the FPGA as a target. A prominent example is loop unrolling.

The compiler-frameworked back end will normally lower the IR into the target’s instruction set and allocate the variables in SSA form to the registers of the CPU. For the FPGA, this compares the translation to an HDL and the allocation of hardware resources (like DSPs). In general, a state machine going through the instructions is created. In the simplest (but also the least efficient) method, this will go through the instructions one by one. To exploit the parallel architecture of the FPGA, it is necessary to schedule the instructions in parallel.

This is typically achieved by viewing the program as a control-data flow graph, while the nodes represent instructions and the edges define the dependencies. The dependencies are then modeled as a set of integer difference constraints \(x - y \leq b\). Further resource and timing constraints are added and finally the system is ready for a linear scheduling objective [4]. The result will assign the appropriate state to each instruction.

Fig. 4: A simplified example schedule, where two independent chains of operations (starting with a load) get scheduled in parallel and later merge during the third operation before the result gets stored in the memory.
One of the greatest advantages of HLS over an HDL is that the development of the algorithm and iterative functional testing can be achieved by emulation on the CPU, which is much faster than simulation or even synthesis. The next step is to optimize based on the reports of the HLS, and only in the last step the performance will be tweaked based on real synthesis results. These reports include loop analysis and area estimates that will guide the following optimization practices.

Besides designing the algorithm in an FPGA-friendly way and structuring the code accordingly, there is a set of annotations to accommodate the source code for the FPGA. These range from annotations to provide additional information or guarantees to the compiler (to help the automatic optimization) up to annotations to instruct the compiler to handle constructs in a certain way. One of the most important candidates is the annotation to unroll loops, as this directly transforms sequential logic (feedback loop) to combinatorial logic. This mainly increases throughput per clock cycle with the trade-off of a higher area demand.

Even though memory accesses will be automatically prefetched, burst-coalesced, or cached (depending on the pattern), it is always advisable to introduce manual buffering. This is preferable in registers (single-cycle access) or otherwise in block RAM (SRAM, multiple cycles but fixed latency). This has the advantage of pipelines not being stalled or it even allowing loops to be unrolled.

To imply registers over block RAM, it is important to access the buffers statically, so it can basically be hardwired instead of an address-based dynamic access. A static access, and hence buffering in local registers, is typically archived by implying a shift-register-like structure, where the new element always gets pushed in at one end of the buffer (fixed location).

To squeeze out the last bit of performance and create a safe area, it is possible to use custom types, like arbitrary precision integers with a custom number of bits or relax floating point operations, so the order can be changed for them to be executed in parallel.

In case it is not feasible to unroll a loop, the compiler will automatically try to pipeline it, for example by overlapping load, computation, and store. Due to loop-carried dependencies, it might not be possible to launch a new loop iteration every clock cycle. This delay is called initiation interval (II). The goal is to minimize the II in order to fully utilize the hardware. This is mainly achieved by introducing local buffers to avoid or at least relax these dependencies.

**Fig. 5**: A simplified pipeline with three stages (columns) shown over time (rows). For example, while the first element A is in the second cycle already in the compute stage, the next element B gets loaded in parallel.
OpenMP Offloading on FPGAs

Even though HLS greatly simplifies the effort of hardware design, it typically only provides an IP block meant to be integrated into the FPGA infrastructure. This still leaves the developer to use the hardware design tools to build a framework for it, e.g. with host (PCI) and memory connections.

The OpenCL toolchains greatly reduce the knowledge needed by providing a fixed infrastructure based on partial reconfiguration, and handling the integration of the user’s component. But, as there is more OpenMP than OpenCL legacy code used within the HPC community and more users are familiar with it, we set out for the quest to bring OpenMP to the FPGA world as part of the ORKA-HPC project [5].

So far, only a limited amount of work was done in that direction. A survey [6] showed that either OpenMP was only handling the configuration of the host side, an old OpenMP version without target pragma was used, or the work only had the character of a prototype and is no longer supported.

Fig. 6: Schematic overview of the ORKA-HPC tool flow to support the OpenMP programming model for FPGAs. (A) The design goal of ORKA-HPC is a fully automated workflow without any required user interaction. The vendor toolchains are used as backends.
A first approach was to utilize the pre-existing OpenCL toolchains [7] of the vendors by outlining the annotated OpenMP regions with Clang, feeding the resulting IR into the OpenCL toolchain, and replacing the host code with OpenCL API calls. This proved to work out but showed limitations regarding the flexibility of the proprietary toolchains.

Therefore, a custom solution was created. It uses the ROSE compiler framework for the outlining task and the standard HLS tools for the generation of the IP blocks representing the user-written C++ functions. These IP blocks are then embedded into a low-level platform created by an ORKA-HPC tool using information provided by the OpenMP annotations. The final implementation of the ORKA-HPC toolchain (see Fig. 6) by integrating the developed individual tools into the overall tool flow is in progress.

(C) The global optimization step realizes a heuristic exploration of design space (clock rate, resource utilization) and gives feedback to the src-to-src compiler (orange lines). The reports of the synthesis, simulation, place, and route tools guide the global optimization for which evolutionary algorithms are used.

(D) The low-level platform and runtime system manages pre-designed low-level platforms. It includes a generic driver and runtime API for the host – FPGA board communication and, for example, the monitoring of platform parameters (temperature, voltages).
References:


SOLVING REAL-WORLD OPTIMIZATION PROBLEMS


MACHINE LEARNING AND BIG DATA


RECONFIGURABLE COMPUTING TODAY


PEER-REVIEWED (154)


PUBLICATIONS


Gunter Mohr, Simon J. Altenburg, Alexander Ulbricht, Philipp Heinrich, Daniel Baum, Chris-tiane Maierhofer, Kai Hilgenberg (2019). In-situ detection in laser powder bed fusion by using thermography and optical tomography – comparison to computed tomography. Metals. (accepted for publication 2019-12-20)


**Publications**


Christine Tawlik, Sabine Limbourg (2019). A bilevel model for network design and pricing based on a level-of-service assessment. Transportation science. (accepted for publication 2019-01-17)


NOT PEER-REVIEWED (10)


BOOKS (2)


DISSERTATIONS (5)


