NSF CC* Planning Project Technical Report

Fostering Data-Driven STEM Research and Education Through Cyberinfrastructure at Illinois State University

Executive Summary

The national research cyberinfrastructure (CI) has become critical to computational and data-intensive research across all of science and engineering (S&E) in the 21st century. However, the concept of campus cyberinfrastructure (CC) is still relatively new to many researchers and institutions. The purpose of this report is to share our experience of planning such a campus cyberinfrastructure (aka science DMZ or simply a data-intensive research network) to foster data-intensive research and education at Illinois State University (ISU).

In this CC* planning project at Illinois State University (ISU), we have established a robust partnership between IT and faculty and identified key science drivers across various disciplines, along with their specific network needs. We have also pinpointed network issues that are affecting, or likely to affect, the productivity of these drivers.

Addressing these challenges, we adopted a systematic approach named Analysis-Architecture-Design, rather than an ad hoc method, to develop a logical, reproducible, and defensible research cyberinfrastructure. This infrastructure, termed the Scalable and Polymorphic Research Infrastructure (SPRI), is designed to meet the diverse and dynamic needs of research and education. SPRI comprises three modular components: the Science DMZ (SDMZ) for high-throughput inter-campus networking, the Campus Research Infrastructure (CRI) for a scalable and dynamic intra-campus network, and the Open Programmable Network Platform (OPNP) to foster research innovation. The modular design principle of SPRI ensures that each component meets its unique requirements while enhancing the overall system performance. This initiative is more than just a network development; it's a vision for the future of data-intensive research at ISU.

The project has significantly impacted the STEM community at ISU, fostering interdisciplinary collaborations and addressing the unique challenges faced by Primarily Undergraduate Institutions (PUIs) like ISU. These challenges include limited resources, diverse research agendas, and a strong focus on education and community engagement. By aligning CI planning with the broad spectrum of research and educational needs, we are integrating our strategic plan to position ISU as a leader in undergraduate education. Furthermore, the project's success extends beyond ISU. Through various dissemination channels like workshops and websites, we are sharing our insights and models with local communities, regional partners, and national institutions, such as those in the Intercollegiate Biomathematics Alliance led by ISU. This project stands as a model CI blueprint, offering a viable solution for PUIs and aiming to enrich the academic experience for underrepresented students across scientific domains.

Key Insights:

- **Collaborative Partnerships**: The foundation of successful CC projects lies in forging robust campus cyberinfrastructure collaborations.

- **Identifying Science Drivers:** Pinpointing science drivers is a nuanced process. It necessitates iterative top-down and bottom-up strategies to convey the essence of CI to researchers spanning diverse fields.

- **Strategic Network Design:** Employing a structured network design methodology, encompassing Analysis, Architecture, and Design phases, proved instrumental in the successful establishment of our campus research network.

1. Background and Motivation

Illinois State University (ISU or Illinois State), the first public university in Illinois, was founded in 1857 as a normal university to prepare the state's teachers. Our institution has a rich heritage as the state's leader in all facets of teacher education, from classroom instruction to educational administration and statewide policy setting. Illinois State has grown its mission over the years and is now a comprehensive university offering more than 160 major/minor options, 42 master's and 10 doctoral programs in six colleges (and expecting a new college of engineering in the near future), but that early emphasis on education continues to shape the University, in particular our values-based commitment to creating an optimal learning environment for all Illinois State students, whether undergraduate or graduate, on campus, or off-campus. Illinois State is recognized as a premier university that has been consistently ranked nationally for its value. Illinois State University received the Elective Classification for Community Engagement from the Carnegie Foundation for the Advancement of Teaching and Learning.

Some may wonder why research is important when Illinois State University has built its reputation as a leader in undergraduate education through outstanding teaching. Research, as identified by the Association of American Colleges & Universities, is an educational practice that has a significant impact on student success. and is core to the University's mission in Educating Illinois. Through research activities, students work one-on-one with a professor, building critical thinking skills while creating new knowledge and learning how to communicate this knowledge to the scholarly community. Such skills are critical for the student's professional development and success in their chosen profession. Currently, ISU has a highly vibrant STEM research community that is inspired by the increasing availability and scales of computation and data. However, ISU's current cyberinfrastructure (CI) cannot accommodate the breadth and depth of their research activities on campus; and cannot allow them to collaborate with a broad spectrum of researchers from peer universities and research institutions.

2. Preparing for research use of network technology on ISU campus

Data-intensive computing opens a new era to research and engineering. As of 2020, the Energy Science network (ESnet), a high-performance network that carries science traffic for the U.S. Department of Energy, is transporting tens of petabytes (PBs) per month, an increase of several orders of magnitude from some years ago. Access to distributed big data, high-performance computing, cloud resources, and other research instruments (e.g., IoT sensors, remote science instruments) has become increasingly crucial and challenging to research and education endeavors in a wide range of disciplines.

While campus IT groups around the nation continue to invest in production-level enterprise-class network infrastructure and services, provisioning very high bandwidth end-to-end connectivity across and beyond campus networks today, amidst firewalls and routing constraints, becomes a major undertaking. Most campus networks today are designed to: 1) serve a large number of users and platforms desktops, laptops, mobile devices, supercomputers, tablets, etc.; 2) support a variety of applications: email, browsing, voice, video, procurement systems, and others; 3) provide security against the multiple threats that result from a large number of applications and platforms; 4) provide various traffic policies that satisfy user expectations. It is not uncommon that most campus networks are not ready to serve large volumes of research and science data transferring with expected performance (e.g., zero packet loss, low latency). Illinois State University is also facing such a challenge as well as an opportunity of transforming STEM research and education.

2.1 Limitations of Current ISU Cyberinfrastructure

Illinois State University maintains connections to several regional Internet Service Providers. The Illinois Century Network (ICN) is supported and maintained by the state of Illinois government, which has a POP location in an on-campus datacenter and provides Internet and DDoS protection to the campus. The Central Illinois Broadband Network (CIRBN) also maintains a POP location in our campus datacenter. CIRBN provides layer 3 Internet connectivity for our campus, but also connects our remote offices to our campus via layer 2. The University of Illinois maintains a research network that also has a POP location in our datacenter. We utilize the Inter-Campus Communication Network (ICCN) for transport to Chicago for peering services with WiscNet and Internet2 (I2). WiscNet is a research and education network in the state of Wisconsin and is managed by the University of Wisconsin. Illinois State University gains membership to I2 through our connection with the Illinois Century Network. ICCN provides transport to Chicago, where we connect with the Metropolitan Research and Education Network (MREN) which peers with I2.

Illinois State University’s campus is robust and stable for day-to-day office and classroom functions but lacks the tuning for sustained high bandwidth data transfers. The existing network supports the bursty nature of Internet destined traffic, e-mail sending, learning management systems, and social media platforms. Data intensive research traffic requires high end-to-end throughput (tens to hundreds of gigabits per second), low latency, high flow capabilities, and deep buffers to support the sustained high bandwidth data transfers. Though the current ISU network has high throughput in its core and datacenter segments, it does not extend to the faculty’s research environment (e.g., labs, offices). The whole ISU STEM research community is currently throttled by 2 Gbps links shared with other daily routine traffic, as highlighted in red in Fig.1.The multi-tier firewall protection is also not designed or optimized for a data-driven research network.

A solution to tackle the challenges above is to reshape campus cyberinfrastructure (CC).

2.2 ISU Research Network

Cyberinfrastructure (CI) combines a spectrum of computing systems, data storage, advanced instrumentation, tools, and services, computational and data analytical skills and expertise, and research communities, all linked by a high-speed network across campus to the outside world. Ideally, an ecosystem of campus cyberinfrastructure should be viewed holistically as an ongoing partnership among the campus research community and central IT organization that is built on a foundation of accountability, funding, planning, and responsiveness to the needs of the community.

To drive innovation, improve research and teaching capabilities and productivity, enhance faculty competitiveness, and foster remote collaborations of resources and people, ISU planned to build a campus research network (aka Science DMZ [1]) environment to provide access to a secured, high-throughput infrastructure for the ISU STEM research and education community, so they will have the ability to access necessary computing and data resources, and collaborate with their peers across the country and around the world.

In this planning project, we designed a high-speed research network for the ISU research community, called ISU Research Network or ISURNet, connecting all identified research groups demanding for high-throughput network service (aka science drivers) across campus, to enable productive data and computing driven research and education. ISURNet is designed to employ agile orchestration applications (e.g., Intent-based Software Defined Networking) to assure holistic high-level network and user management with assured end-to-end network performance.

This designed research network will not only allow consolidation of distributed data repositories and advanced computing resources across campus, but also improve the speed and Quality of Service (QoS) when connecting to the national cyberinfrastructure ecosystem, e.g., FABRIC [3], Chameleon [6], CloudLab [7], Jetstream [8], XSEDE [11], Open Science Grid (OSG) [12], Pacific Research Platform (PRP) [13], and GENI [14]. Our goal in ISURNet is to couple all relevant research groups identified in this project at 10+ Gb/s backbone, field upgradeable to 100 Gb/s, that will bring each identified research group an average increase in bandwidth between 10 and 80 Gb/s.

The three objectives of this planning project include: 1) Increasing awareness of campus cyberinfrastructure across campus and establishing campus cyberinfrastructure partnerships between faculty and IT leadership; 2) Identifying and analyzing science drivers across campus; 3) applying a systematic approach to design a *logical*, *flexible*, *extensible*, *reproducible*, and *defensible* Campus research network to foster data-driven STEM research and education.

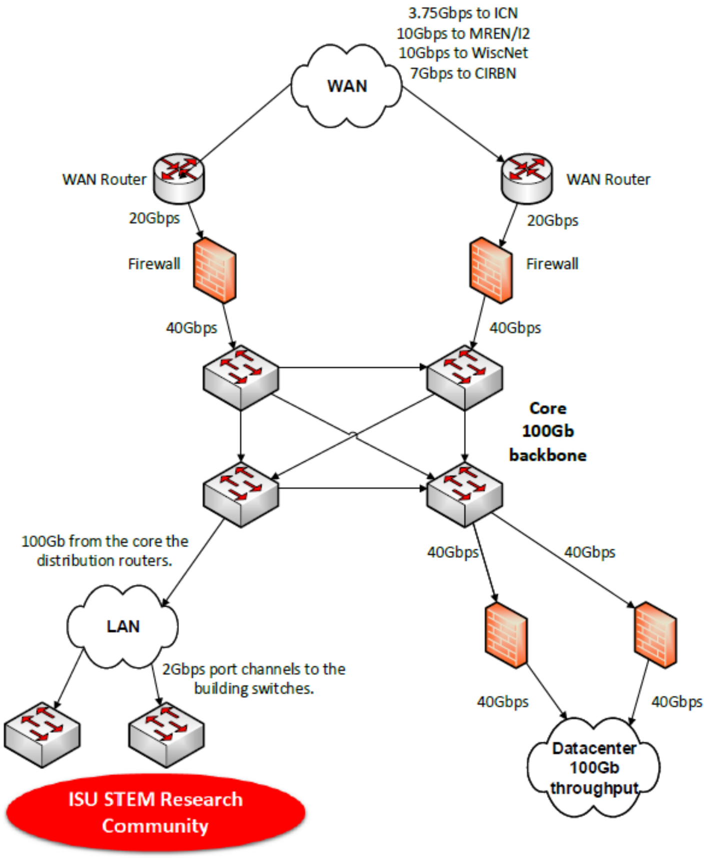

Fig. 1: Current ISU Network Diagram

3. Partnership between Faculty and IT Leadership

This endeavor of this planning project aligns with the latest ISU Strategic Plan [2], in which ISU encourages enhancing organizational infrastructure to support innovation and collaboration, involving more faculty, staff and students in research opportunities locally, regionally, and globally. With a data-intensive research network, ISU could exponentially accelerate the inspiring and cutting-edge research being done by faculty to advance research and education.

Writing a Campus CI plan is not the most challenging part, implementing it is harder. Having a Campus CI plan is only useful if the campus is committed and there is full support from the university senior administration. Fortunately, we have this support at ISU.

Vision and Strategy: ISU University Office of Technology Solutions (OTS), under the leadership by Associate Vice President and Chief Information Officer Charles Edamala, co-PI of this project, has made cyberinfrastructure a strategic priority of the university. Two of co-PI Edamala's strategic priorities for the university include building "a comprehensive and integrated data environment" and "advanced research computing." Both of these require a low-latency high performance programmable research network infrastructure that allows easy movement of data across campus, between research groups (internal and external), data stores, and to advanced computing resources. Funding for this planning project further allowed the PIs create a comprehensive blueprint for improving ISU CI, consolidate the CI ecosystem on campus, and keeping ISU competitive in providing premier undergraduate and distinguished graduate programs in the state of Illinois.

Formality: To better serve the ISU research community and the growing demands for campus cyberinfrastructure, a new *Research Computing Advisory Council* (RCAC) was formed in 2018, reporting directly to the Office of Technology Solutions (OTS). RCAC acts as a bridge between the University's research faculty and information technology professionals. A university-wide initiative has also been launched to develop and implement a bold strategic vision for centralizing the advanced research computing and data cyberinfrastructure ecosystem at ISU. RCAC provides a means for efficient and effective communication, deployment, training, and collaboration aimed at increasing the research capabilities and productivity through not only the planned *ISURNet*, but to the entire ISU CI ecosystem. Working together across all ISU units, RCAC provides the university community with strategic advanced CI leadership, coordinated investments in CI and related expertise, and nurtured CI-enabled multidisciplinary research. All PIs are the board members of RCAC.

The Infrastructure, Operation, and Networking Division (ION) within OTS is responsible for the design, implementation, operation, maintenance, and evolution of current CI at ISU, including central voice, television, wireless, and data networks. These high performing networks are critical utilities at the university and are fundamental to the success of research, teaching, and learning. The goal of ION, aligning with ISU's vision [2], is to not only provide seamless and pervasive networking for the university community, including production level enterprise-class high-speed network infrastructure and services, but also to offer services that give researchers a competitive edge. Participating in and supporting projects similar to what we are planning creates a transformative environment where campus IT, and in particular ION, becomes seen as a strategic asset rather than a utility provider. ION's role in this CC planning project will be assisting the design and prototyping of *ISURNet*, including perfSONAR, end-to-end performance testing, and IPv6 deployment.

Partnership: CC Partnership is not just about vision, strategy, and formality. Instead, a real CC partnership must be established over time among involved individual researchers, IT professionals, and university administration via continuous and consistent effective communications to share visions about CC among individuals, and eventually build mutual trust on the agreed commitment and responsibilities.

During this 16-month planning project period, tens of meetings, seminars, and workshops have been conducted at different levels with various focused CC-related topics. We just listed a few below as examples:

- A university-level presentation was given to the ISU president and most (if not all) deans to elaborate on the significance and impact of this planning and other potential CC-related projects.
- Multiple presentations and discussions were conducted between the PIs and department chairs, directors of various research centers, and research lab leaders.
- Multiple one-to-one meetings have been conducted between the PIs and individual researchers.
- Multiple meetings have been conducted between the PIs and IT engineering teams.
- Several meetings have been conducted between the PIs, IT team, and various vendors.
- A continuous CC project team has been formed and bi-weekly project meetings are conducted to prepare the next move after this planning project.

4. Identifying Science Drivers at ISU

In the context of the National Science Foundation (NSF) campus cyberinfrastructure programs, "science drivers" refer to the specific research needs and scientific challenges that drive the development and implementation of cyberinfrastructure. These drivers help shape the technology, tools, and services required to support cutting-edge scientific research.

It will surprise you how many science drivers possibly exist on your campus. We had such a pleasant experience at ISU. However, it is also worth noting that identifying science drivers is not a trivial task at all. We are talking about identifying potential data-intensive researchers from hundreds or even thousands of faculty members from various disciplines in your institution.

The experience we can share is to use both top-down and bottom-up approaches to accomplish this task. More specifically, in the top-down approach, try to explain the concept of science driver to department chairs and even college deans, and ask if they know anyone from their units you may further contact; in the bottom-up approach, start chatting with faculty you know and ask them to bring you more names, and then grow the list and go verify if they are science drivers for your CC projects. Furthermore, your campus media (websites, various reports, social media, campus journals, etc.) may provide useful clues to complement your search.

In this section, we summarize the identified science drivers. Each of the projects is led by faculty members in the academic departments. They have identified at least one of the three network issues, which are or soon will be impacting their research productivity.

4.1 Quantum-Mechanical Approach to the Laser-Assisted Vacuum Decay

Lead: Q. Charles Su and Rainer Grobe, Distinguished Professors of Physics

Su and Grobe, both distinguished professors of physics at ISU, conducts research on the interaction of laser light with matter. The groundbreaking, theoretical work of the duo earned them the designation of Cottrell Science Scholars, and the Research Corporation for Science Advancement (RCSA) named them among the top physics scholars in the country. There are only a few centers around the world engaged in research similar to the work of Su and Grobe that focus on the interaction between extremely intense laser light with atoms. Their research has been supported by grants from NSF (e.g., Award # 0758058, #0456790, #2106585), the Department of Energy, and the Research Corporation.

Simulating the interaction of a single atom with a laser pulse over a picosecond ($10^{–12}$ seconds) timeframe with a temporal resolution of a femtosecond (10^{-15} seconds) can generate millions of data points. If each data point is represented using 64 bits (8 bytes), even a single simulation can produce 1~10 GB data. If each process simulates a different atomic interaction and generates gigabytes of data, the total data generated can quickly reach 100 TB. These data need to be processed in their specific simulation workstation in their research lab and transferred back and forth to the ISU HPC system for specific data manipulations or computation, and thus **throttled by 2 Gbps links**.

Their research on diagnosing laser fusion reactions or studying harmonic light generation can involve capturing high-resolution images or spectra. A single high-resolution image can be several megabytes in size. Capturing thousands of such images during an experiment can result in data storage needs in the 10~200 TB range. These images are also needed to be transferred between the simulation workstations and HPC.

Their research group began to employ machine learning techniques to help derive physical laws or predict physics outcomes in regions where it was considered not possible to reach. Training a deep neural network on a dataset derived from quantum simulations can generate hundreds of gigabytes or even terabytes of data in each simulation. The training process itself, especially when using techniques like symbolic regression based on evolutionary algorithms, can produce intermediate datasets that are several times larger than the original. These data again need to be exchanged between their simulation workstations and HPC or storage servers in the ISU data center.

Collaborative research (e.g., with Fermi National Lab or Fermilab) often involves real-time data sharing and analysis. With frequently cited publications [21, 22, 23, 24, 44, 45, 46, 47], the associated datasets, supplementary materials, and simulation results need to be shared with the global research community, which is nearly impossible for datasets at TB level now.

Workflow & Requirements: 1) Various datasets ranging from 10 GB to 200 TB in different research activities (e.g., High-Resolution Quantum Simulations) need to be exchanged between simulation workstations and HPC (or storage server) necessitates 20+ Gbps LAN links; 2) Real-time data analysis with external collaborators (e.g., Fermilab) requiring rapid synchronization and 10~100 GB data exchange in each run necessitates 10+ Gbps WAN links; 3) Data sharing (up to hundreds of TB) to the research community necessitate 10+ Gbps WAN links.

4.2 Intercollegiate Biomathematics Alliance (IBA)

Lead: Olcay Akman, Professor of Mathematics, Director of IBA, Co-PI

The Intercollegiate Biomathematics Alliance (IBA) is a collaborative consortium of institutions ranging from R1 research institutions (e.g., George Mason University) to R2 PUIs like ISU, dedicated to advancing biomathematics and bioinformatics research and education. IBA promotes innovative research, addressing complex biological problems through mathematical and computational approaches. Simultaneously, IBA plays a pivotal role in enhancing education by offering specialized courses, workshops, and conferences (e.g., NSF Award #1332395; #1649061, #2318936).

One pressing IBA research area is the modeling of disease propagation and their social impacts. In their study of the intertwined dynamics of the COVID-19 pandemic and social epidemics, they utilized multifaceted models that integrated real-time data from COVID Tracking Project, a volunteer organization providing one of the most comprehensive datasets on COVID-19. With current ISU network infrastructure, such research projects are very difficult if not impossible. This research harnessed datasets and simulation results among 1 ~ 10 TB, which also need to be shared to other IBA institutions and selected for various graduate projects but also bottlenecked by 2 Gbps and congested WAN links.

IBA's analytical investigation into the efficacy of N95 respirators for the public was anchored via their novel math modeling. This project synthesized experimental data from global respiratory studies and health trials, amassing a dataset of approximately 5+ TB. The data, sourced from international health research institutions, such as CDC, National Institute for Occupational Safety and Health (NIOSH), and Harvard Dataverse, was pivotal in offering a mathematical perspective on public health recommendations. These types of studies again suffered from degraded data transfer performance.

In addition to research, the IBA is deeply committed to education in biomathematics. IBA offers specialized courses that incorporate big data analytics, machine learning, and simulation-based methods. These courses are designed to be hands-on and are hosted on cloud-based platforms located at the ISU data center, requiring a stable, high-throughput network with a minimum bandwidth of 5+ Gbps. Each course session is expected to generate and consume up to 100 GB of data, including student submissions, simulation results, and video lectures. All these activities are currently limited.

Workflow & Requirements: 1) Disease Propagation Modeling (e.g., disease spread like COVID-19) at IBA, leveraging interdisciplinary expertise, and the data sources globally distributed, necessitates a 10+ Gbps network to handle real-time data from diverse sources and manage 1-5 TB datasets per experiment; 2) IBA's hands-on courses in biomathematics, hosted on cloud platforms, integrate big data and simulations, demanding a stable 5+ Gbps network to manage up to 100 GB of data per session, including lectures and student work; 3) IBA's national collaborations and online events, involving real-time data sharing and analysis, and live-streaming, require a combined throughput of 10 Gbps, handling up to 2.2 TB of research data and high-quality event broadcasts.

4.3 Computational Biochemistry

Lead: George L. Barnes: Professor of Chemistry, Department Chair, Co-PI

The research endeavor led by Co-PI Barnes represents a highly impactful contribution to the field of tandem mass spectrometry (MS2) and collision-induced dissociation (CID) [54, 59, 61, 62]. Research within MS2 offers substantive applications in proteomics [48, 52, 53], metabolomics [60], and forensic sciences [63]. The focal point of the research is the computational simulation of post-translational modifications (PTMs) in peptides [49, 50, 51], a subject matter with profound implications for targeted therapeutics and molecular identification methodologies.

Research in the Barnes Group relies on high-throughput networking and high-performance computing along with strategically located storage to all for large-scale data management. The computational architecture underpinning this research [55, 56, 57, 58] is non-trivial and necessitates a multi-faceted technological infrastructure. Their projects employ HPC to execute direct dynamics simulations, each of which involves several thousand to tens of thousands of individual runs. These simulations are particularly computationally intensive when the research objectives involve the elucidation of rare stochastic events within the simulation parameters. Upon the completion of these simulations, the data undergoes rigorous post-simulation analysis employing graph theory-based methodologies. This involves a frame-by-frame dissection of each trajectory's simulation data to ascertain alterations in molecular connectivity over time. Subsequently, these key structures are characterized at a higher computational level through Density Functional Theory (DFT)-based calculations, specifically utilizing the ωB97XD/aug-cc-pVTZ level of theory and basis set. This computational rigor enables the research team to delve into the minutiae of chemical reactions, offering insights that are unattainable through conventional methods.

Data management is another critical component for their research. Most their projects are projected to generate at least 1 TB of new data sets annually. This data is not merely archived; it is strategically located in proximity to the HPC resources to optimize data retrieval times, thereby enhancing computational efficiency. However, a high-throughput network infrastructure is essential to ensure the seamless transfer of large data sets between the Co-PI's workspace and the data storage servers, a necessity for tasks such as real-time trajectory visualization, which is only facilitated and thus throttled by 2G bps links. The growing datasets need to be shared in their collaborations, which is difficult now.

Workflow & Requirements: 1) Barnes' research in MS2 and CID, pivotal for proteomics and metabolomics, of direct dynamics simulations, followed by graph theory analysis and DFT calculations at the ωB97XD/aug-cc-pVTZ level, generating insights into intricate chemical reactions, all hinges on high-throughput networking between HPC and data storage, necessitating 25+ Gbps internal throughput. 2) Annually producing 1 TB of data, efficient storage near HPC resources is crucial, and a 2Gbps network is vital for swift data transfers and real-time trajectory visualization, underscoring the need for a high-speed network.

4.4 Intelligent Network Operations (AIOps)

Lead Yongning Tang: Professor of Computer Networking, PI

Modern computer networks are intricate, spanning data centers, network infrastructure, and IoT devices, which lies a challenge: How do we manage such complexity? The answer is AIOps. By integrating AI and ML into network management, AIOps offers a transformative, data-driven approach that transcends the limitations of traditional methods. The research group led by PI Tang has been applying AIOps to various network management areas, from performance to security, from traditional networks to SDN, from wired backbone infrastructure to 5G and IoT.

In the research project of detecting malfunctioning IoT devices and network outages in real-time, multiple big datasets have been used, including ~100GB daily device telemetry data, ~150 GB network traffic logs, ~1TB device error logs, and about 10TB historical device data. Such huge datasets need to be retrieved from a storage server and computed in our dedicated machine learning servers via a 2G bps campus link currently. This is a joint research project with external researchers from Liberty University with data from their testbeds, which is hard to retrieve.

In the research project of creating a scalable service that can analyze network traffic patterns and promptly detect anomalies, high volume of network flow data is imperative for the designed machine learning model to learn intrinsic network behaviors. For example, a single network flow record might be around 250 bytes. The data collected from a medium-sized network with 100,000 flows per second, could accumulate to 2.16 TB daily. A machine learning epoch based on 10-day historic data as a minimum to learning the network behavior requires 20 TB data consistently fed to the machine learning server, currently via 2 Gbps links.

In the project called Intelligence Enabled SDN Fault Localization via Programmable In-band Network Telemetry (INT), it aims to enhance Software-Defined Networking (SDN) by introducing intelligent fault localization using programmable in-band network telemetry. In-band network telemetry involves the collection of data directly within the data packets as they traverse the network. In a high-speed SDN environment with a throughput of 10 Gbps, if we assume that 1% of the data is telemetry data, we're dealing with 100 Mbps of telemetry data. Over a day, this amounts to approximately 1.08 TB of telemetry data. Analyzing this data to localize faults in real-time is a data-intensive task, necessitating the capabilities of AIOps.

Workflow & Requirements: 1) Our research in AIOps is data-intensive, necessitating a 10 Gbps intra-campus network to handle datasets between 200 GB to 1 TB for each machine learning epoch. 2) Data sharing with external collaborators such as IoT data collected from the testbeds hosted on each end necessitate a 10 Gbps inter-campus connection for effective collaboration. 3) SDN and INT/p4 related research projects need a real campus infrastructure level open environment for transferring related scientific data flows to emulate and study different research problems.

4.5 Center for Cybersecurity Research and Education (CCRE)

Lead: Dmitry Zhdanov, Associate Professor of Cybersecurity, Director of CCRE

Our cybersecurity research team aims to revolutionize the field of cybersecurity through real-time threat intelligence and analysis. This objective is particularly data-intensive, requiring a high-throughput network with a minimum of 10 Gbps for real-time data streaming but currently throttled by 2G bps links. We anticipate that daily datasets for this research will range from 100 GB to 10 TB.

Another critical area of focus is advanced network forensics, which involves the analysis of large network traffic pcap files. The nature of this research necessitates a high-speed data transfer capability, ideally around 10 Gbps or above. A single investigation under this objective may generate up to 1 TB of pcap files.

Furthermore, our faculty conduct research on DDoS attack simulation and mitigation. The nature of this type of research is highly data-intensive and requires a robust network infrastructure capable of handling 40 Gbps, and thus limited by 2G bps links again. We estimate that each simulation will generate up to 500 GB of log and traffic data.

On the educational front, the cybersecurity faculty are committed to providing virtual cybersecurity labs for skill development. These labs require seamless operation of virtual machines, necessitating a high-speed network access of at least 4 Gbps, and thus limited by the shared WAN link. Each lab session with 20~25 students is expected to generate up to 200 GB of data. Additionally, we offer remote access from community colleges and high schools to advanced security tools for educational purposes. Smooth operation of these tools is facilitated by high-speed networks with a minimum throughput of 500 Mbps. Daily usage for this educational activity may result in up to 50 GB of data. We hold annual Central Illinois High School Cyber Defense Competition for trained high school students since 2012, and GenCyber summer camp since 2020, sponsored by National Security Agency (NSA) and NSF.

Workflow & Requirements: 1) The cybersecurity team's real-time threat intelligence research is data-intensive, requiring a 10 Gbps network for daily data streams ranging from 100 GB to 10 TB; 2) DDoS attack simulations are exceptionally data-heavy, necessitating a robust 40 Gbps network to manage up to 500 GB of log and traffic data per simulation, and virtual cybersecurity labs for education need 2 Gbps for seamless VM operation, generating up to 100 GB data per session; 3) Remote access to advanced security tools for educating the local community requires 4 Gbps network.

4.6 Unveiling the Mysteries of the Universe Through High-Throughput Computing

Lead: Danial Holland, Professor of Physics, Dept Chair; Matt Caplan, Assistant Professor of Physics

This research group employs MESA, the benchmark open source 1D code for stellar evolution, to explore stellar evolution influenced by low mass central black holes. This research aims to understand quasi-stars in the early universe and potentially constrain primordial black holes as dark matter candidates. Individual models, representing single star evolution sequences, are computed within a day on a single CPU, generating data from 50~100 GB. Comprehensive 'grids' of these models, which survey the evolution across varied parameters, can produce 100 GB ~ 10 TB. Given the intricate nature of these models, data is frequently transferred from the HPC to the servers in their research labs for in-depth data analysis and visualization, which is currently throttles by 2G bps.

The computational models are not only complex but also data-intensive, demanding real-time data from telescopic observations (e.g., SDSS, Pan-STARRS), particle accelerators (e.g., LHC at CERN, Fermilab), and astrophysical databases (e.g., NASA ADS, NVO, SIMBAD). A robust network infrastructure, with a bandwidth of at least 20 Gbps, is crucial to manage the 10 to 20 TB datasets each astromaterial simulation requires. The dark matter research is similarly data-heavy, with simulations necessitating a 10~20 Gbps network to handle up to 15 TB of data per simulation. This project also delves into atomistic simulations of neutron star interiors, which are semi-classical in nature. These simulations, akin to liquid crystal structures, involve up to 102,400 nucleons and generate TBs of data, necessitating efficient data transfers for visualization. All external data retrievals are currently highly inefficient if not infeasible to be conducted at ISU.

The research team also has many collaborative projects, involving real-time data sharing with external scientists, further require a 20 Gbps network, producing up to 10-15 TB of combined research data. For instance, collaborations with renowned scientists like Yuri Levin (Columbia) and Katerina Chatziioannou (Caltech) are funded by the Simons Collaboration on Extreme Electrodynamics in Compact Sources. The research on Coulomb plasmas, large scale N-body problems, collaborating with Dr. Simon Blouin (University of Victoria) and Dr. Evan Bauer (Harvard-Smithsonian Center for Astrophysics), along with ISU physics graduate student Dany Yaacoub actively involved in running related simulations. Such extensive collaborative research activities are difficult to conduct at ISU.

Workflow & Requirements: 1) The research projects such as Coulomb plasma simulations, Neutron star interior simulations, and MESA stellar evolution modeling generating datasets up to 20+ TB per study, which necessitates a 10+ Gbps intra-campus (between HPC and Workstations) network throughput; 2) The research requiring for retrieving large (~10+ TB) external data from public repositories (e.g., SDSS, Pan-STARRS, LHC) necessitates a 10+ Gbps network throughput to WAN; 3) Extensive research collaboration, especially when real-time analysis of large datasets (1~10 TB) is required, with external scientists (e.g., Columbia, Caltech) necessitates a 10 Gbps network throughput to WAN.

4.7 CubeSat and 5G research group

Lead: Will Lewis, Asst. Prof. of Information Systems; Sumesh Philip, Asso. Prof. of Cybersecurity

CubeSats are standardized, miniaturized satellites that are designed to be modular and easily deployed in low-earth orbit (LEO) for various missions. including mobile networks (5G/6G), global Internet connectivity, disaster response communications, precision agriculture, and the Internet of Things (IoT) by offering a satellite-based infrastructure for improved communication and data gathering.

The research team currently working on several CubeSat and IoT related projects. The team installed a CubeSat ground station at ISU two years ago, collaborating with Fermi National Lab (Fermilab) and the Laboratory for Advanced Space Systems at the University of Illinois at Urbana-Champaign (UIUC) on a CubeSat mission named DarkNESS to search for evidence of dark matter by looking for signs of the decay of sterile neutrinos at the center of our Milky Way galaxy. For such a mission, a set of CubeSats are equipped with a range of specialized instruments to detect and analyze high-energy particles and radiation, such as X-ray Detectors, Gamma-ray Detectors, Particle Detectors, and Spectrometers. Due to the nature of the project, high-resolution spectral data is continuously received from visible CubeSats and shared with the external collaborators for real-time image processing and data analysis. The daily collected data size collected from each ground station is between 10~100 GB. But more crucially, the data need to be shared to the external collaborators for real-time data aggregation and analysis; but throttled by 2G bps links.

CubeSats are constrained by limited communication bandwidth and data rates, which are further exacerbated by interference from various sources. The ability to predict and mitigate interference is crucial for optimizing data rates in CubeSat Software-Defined Radios (SDRs). The research team is developing a predictive interference mitigation system for CubeSat SDRs to optimize data rates. SDN can be easily integrated with Software-Defined Radios (SDRs) on CubeSats, allowing for a more cohesive communication strategy, where the network and the radio are in sync, adapting to real-time conditions. However, due to the lack of campus level SDN infrastructure, these types of projects are restricted to simulations conducted in a research lab.

A cross-disciplinary project called "VisionWalk" aims to help vision impaired students free walk on a campus environment. The project relies on the Ultra-reliable Low Latency Communication (URLLC): provided by the ISU private 5G network, and centralized network control from SDN. For vision-impaired students, real-time video feed processing is crucial. Any delay in processing and relaying information can pose safety risks. 5G ensures that data from cameras and sensors is transmitted with minimal delay, allowing for almost instantaneous feedback. SDN, from another aspect, allows dynamic path selection based on real-time network conditions. If a particular path experiences congestion, SDN can reroute the data to ensure consistent high throughput. Considering the real-time video feed from cameras, sensor data from wearable IoT devices, voice commands and backend data processing and machine learning models, the total throughput is approximately 10 Gbps and the total daily generated data size is about 100 TB depending on the tested nodes. This type of projects completely relies on an open programmable SDN platform, which can only be emulated now with high limitations. The dataset also needs to be shared between ISU and its collaborators in OSF hospital, and thus currently limited.

Workflow & Requirements: 1) External collaborations on CubeSat DarkNESS projects require a stable 2+ Gbps WAN connection for timely data sharing. 2) SDR needs to be seamlessly integrated with SDN to provide adaptive communication and interference mitigation. 3) 5G as an enabler for delay sensitive research projects can be significantly enhanced via a SDN open platform.

4.8 Natural Language Processing (NLP) and Understanding (NLU)

Lead Xing Fang, Associate Professors of Computer Science

Natural Language Processing and Natural Language Understanding rely quite heavily on Language Models (LMs). These models are required to learn distributions over large amounts of data to encode the nuances of natural language and capture the intricate aspects of human language, ranging from semantics and syntax to morphology. Beyond these linguistic features, LMs also encapsulate psycholinguistic and sociolinguistic dimensions, making them indispensable for a wide array of NLP applications such as question-answering, machine translation, dialogue systems, and automated summarization.

Given the complexity of language, the training of these models demands extensive datasets. Here are examples of some specific datasets and their sizes that our research will utilize: 1) CommonCrawl (Common Crawl Corpus): 50-90TB - General-purpose web text data. With the current average speed of the Internet (1Gbit/sec) can take anywhere from 130 to 200 hours to download only one month of CommonCrawl; 2) SQuAD 2.0: 2GB - For question-answering systems; 3) WMT19: 10GB - For machine translation tasks; 4) Conversational Intelligence Challenge (ConvAI): 5GB - For dialogue systems and conversational agents; 5) CNN/Daily Mail: 4GB - For text summarization; 6) CodeSearchNet: 20GB - For code-related language tasks; OpenSubtitles: 20GB - Subtitle-based translation and multilingual tasks.

Given the data-intensive nature of this research, aggregating to well over 50TB, the need for high-throughput networking is imperative but currently limited. A dedicated Data Transfer Node (DTN) would significantly accelerate the data transfer rates, thereby enabling our researchers to initiate and execute experiments more efficiently. This is particularly crucial when dealing with real-time applications like dialogue systems or time-sensitive tasks like live translation.

Workflow & Requirements: Machine learning projects specifically large language model training heavily rely on obtaining sufficient relevant data, which necessitate high speed (10+ Gbps) networks for intra- and inter-campus data transferring (e.g., for 50-90 TB datasets).

4.9 Water and Remote Sensing Research

Lead: Eric Peterson, Interim Chair of Environment Science Dept, University Professor of Geology

This research team focuses on a myriad of projects that study the water cycle, climate, and human impact on a global scale. The lab employs a range of tools, including physically-based hydrologic models, machine learning, and large data analysis, particularly satellite data. The lab's computational capabilities are robust, featuring high-performance computers, large-capacity data servers, and a multi-level GIS and remote sensing data storage system. Additionally, the lab utilizes Unmanned Aircraft Systems (UAS) for hydrology and environmental monitoring.

Given the data-intensive nature of these projects, a high-throughput network infrastructure is indispensable. For instance, the lab's work on quantifying the effects of climate change on water resources like groundwater, lakes, and reservoirs requires real-time data retrieval from various sources such as NASA's GIOVANNI and the United States Geological Survey (USGS). These datasets are often voluminous, with a single project potentially generating up to 15 TB of data. A network throughput of at least 40 Gbps is essential to handle such large data volumes efficiently, but currently limited.

Similarly, the lab's UAS-based research for hydrology and environmental monitoring involves the acquisition, processing, and analysis of high-resolution spatial data. These activities generate datasets received through ISU 5G private network that can range from 10 to 20 TB, depending on the scale of the study area and the resolution of the sensors used. A high-throughput network with a minimum bandwidth of 35 Gbps is crucial for the seamless transfer and processing of these large datasets, but currently limited by the limited campus infrastructure. Furthermore, SDN as a mechanism to orchestrate the 5G functionalities to support this project is currently not available.

Moreover, the lab's international collaborations necessitate real-time data sharing, further emphasizing the need for a robust, high-throughput network infrastructure. A single international collaborative project can generate up to 25 TB of research data, requiring a network throughput of at least 20 Gbps for efficient data exchange, which is current unavailable.

Workflow & Requirements: 1) GIS modeling especially fed by data collected via UAS can generate huge amount of various sensory data (20+ TB daily), necessitating a 40 Gbps dedicated campus network; 2) External collaborations based on large datasets sharing necessitate at least 20 Gbps WAN connection. 3) 5G/SDN as an enabler and catalyst for UAS based projects are imperative but currently limited by emulations.

Table 1: Summary of Science Drivers and Their Network Requirements

Science Driver	Representative Activities	Data Transfer Load	Network Requirements
Intense Laser Physics	High-Resolution Quantum Simulations; Collaboration & Data Sharing	10 GB ~ 200 TB between HPC & department; 100 TB to external	10+ Gbps Intra-campus throughput; 10+ Gbps WAN throughput.
Biomath & Bioinformatics	Disease Propagation Modeling; IBA's national collaborations	1~10 TB datasets from external to HPC & department	10 Gbps Intra-campus throughput; 15+ Gbps WAN throughput.
Computational Biochemistry	Direct dynamics simulations, followed by graph theory analysis	100 GB datasets frequently exchanged between HPC and storage; 1 TB between Storage and department	10+ Gbps Intra-datacenter throughput; 2+ Gbps Intra-campus throughput.
Intelligent Network Operation (AIOps)	IoT, SDN, In-band telemetry, Blockchain	1-20 TB datasets between ML server & Storage; 1-10 TB to external; 10-20 TB over SDN	10 Gbps Intra-campus throughput; 10+ Gbps WAN throughput; 10+ Gbps open SDN platform
Cybersecurity Research	Threat Intelligence; DDoS; Virtual labs	0.11-10 TB between Storage to department; 200 GB – 1 TB to and from external	10-20 Gbps Intra-campus throughput; 4+ Gbps WAN throughput;
Astrophysics & Astromaterials	Coulomb plasma simulations, Neutron star interior simulations, stellar evolution modeling	1-20 TB between HPC & department; 10 TB to external	10+ Gbps Intra-campus throughput; 10+ Gbps WAN throughput.
CubeSats & 5G & IoT	DarkNESS; SDR/SDN; VisionWalk	1-100 TB between storage and department; 10~100 TB to external; 10+ TB over SDN	10+ Gbps Intra-campus throughput; 2+ Gbps WAN throughput; 20+ Gbps open SDN platform
Natural Language Processing	Language Modeling; automated summarization.	50-90 TB between storage to ML servers; 50-90 TB from external	10+ Gbps Intra-campus throughput; 10+ Gbps WAN throughput;
Water & Environment	GIS modeling; UAS sensing; Collboration	10-35 TB between storage and department; 10~15 TB to external; 10+ TB over SDN	10+ Gbps Intra-campus throughput; 10+ Gbps WAN throughput; 15+ Gbps open SDN platform

4.10 Science Driver Analysis: Water and Remote Sensing Research

In the following, we elaborate on one example below to take a closer look at what a science driver entails.

Prof. Wondwosen Seyoum leads the Water and Remote Sensing Research (WRES) Lab and studies the water cycle, climate, and human impact all around the Globe at ISU. They measure, characterize, and simulate processes in the water cycle using field observations, modeling, and remote sensing. The goal of the WRES lab is to apply the knowledge to sustainable future water management of threatened global freshwater resources due to human impacts and global change. The lab is working on several projects: 1) Quantifying and exploring the effect of future climate change and human impacts on water resources, such as groundwater, lakes, and reservoirs; 2) Assessing the sustainability of threatened aquifer systems; 3) Evaluating the impact of drought and climate variability on water availability; 3) Integrating remote sensing techniques in hydrology in ungauged basins; 4) Understanding groundwater feedback to the land-atmospheric processes and climate; 5) Modeling the effect of agricultural management practices on water pollution, fate, and transport of contaminants.

In the WRES Lab, they use various tools to simulate and investigate the terrestrial hydrological cycle, including physically based hydrologic models, machine learning, and large data analysis such as satellite data. Their computational lab is equipped with high-performance computers for hydrogeologic and water resources modeling, spatial analysis, and remote sensing. In addition to the workstations, they have large-capacity data servers with 64 GB RAM and spacious disk space storage of up to 24 TB. They have a multi-level and organized in-situ GIS, and remote sensing data storage system. Their Unmanned Aircraft System (UAS) - remote sensing research supported by DJI M200 V2 and Phantom 4D drones and sensors including FLIR radiometric thermal, Zenmuse X4S – RGB, and Micasense RedEdge-MX multispectral sensors. The followings are concrete examples of their ongoing and past research projects:

- Unmanned Aircraft Systems (UAS) for Remote Sensing In WRES lab, they use UAS for hydrology and environmental monitoring. Accordingly, they are developing acquisition, processing, and analysis of UAS data for water cycle monitoring, including discharge, water inundation, water quality, soil moisture, and snow/ice. They are also testing and developing methods for agricultural applications such as plant phenology analysis, plant health, and water stress studies. Further, they are developing UAS capability for subsurface characterization such as groundwater, geological, and agricultural (e.g., tile drainage mapping) applications. Evaluation and integration of UAS data with in-situ observation and hydrologic models is their target to quantify the water cycle and improve predictive capacity at various scales, from field to watershed scale.
- Past, present, and future climatic impact on man-made and natural reservoirs Man-made and natural reservoirs are essential to water resources. Reservoirs in small-scale watersheds have limited areal extent and highly rely on the intensity and frequency of precipitation. As a result, storage in these reservoirs is prone to climatic variability, which will potentially be exacerbated by future climate change. Understanding how these systems are responding to short- and long-term climate variability as well as to future climate conditions is essential for a sustainable water future.
- Climate sensitivity to shallow subsurface water Dynamics Land, ocean, and atmospheric processes interact with each other and potentially impact the climate in various ways. Ocean-atmosphere interactions such as those produced through sea surface temperature variations, for instance, often impact meteorological conditions that may enhance extreme precipitation and drought events. This project explores the relationship between hydrological changes related to shallow subsurface waters and how those fluctuations may relate to and impact climate.

- Satellite applications in hydrology Most basins throughout the world, specifically those in non-industrialized countries, are poorly gauged, as a result prediction of the hydrologic cycle is challenging. Fortunately, recent advancements in satellite-based hydrology have demonstrated that some water cycle components can be directly or indirectly estimated from space. In this theme, the group is working on enhancing the utility of satellite products (e.g., GRACE, TRMM, LANDSAT) in monitoring the water cycle, by merging satellite products with machine learning models.
- The effect of cover crops on nutrients load reduction Nutrients loading into streams and rivers from agricultural runoff is a major concern in the Mississippi River Basin: (1) reducing the fertility of soil and (2) deteriorating the quality of streams and rivers, including eutrophying surface water bodies, and creating water quality problems in watersheds downstream. Various best management practices (BMP) are being implemented to reduce nutrient loading in the Mid-west region. The research team is collaborating with research groups across campus on a project aimed at characterizing and simulating the effect of management practices in nutrient loss reduction.

Many research projects in the WRES lab are becoming increasingly data-intensive, requiring robust data storage, management, and analysis capabilities. Many of those datasets are high volume and need to be retrieved regularly from various data sources (e.g., NASA, USCS), as elaborated below.

- Most of the satellite-based hydrology data (e.g., TRMM precipitation, NLDAS land surface data) can be accessed and analyzed from GIOVANNI, which is a NASA-maintained application that allows you to visualize selected geophysical parameters in real-time.
- Most of the remote sensing land data can be accessed via USCS (United States Geological Survey) maintained data rsource called LP DAAC
- GRACE terrestrial water storage anomaly data can be obtained from GRACE Tellus. GLDAS monthly water content for GRACE. GLDAS is Land Water Content (monthly) Data from the Noah land hydrology model in the Global Land Data Assimilation System (GLDAS), and need to be retrieved dynamically for its accuracy.
- Interactive GRACE data visualizations – CU GRACE Data Portal, CNES/GRGS plotter
- Meteorological data (e.g., precipitation, air temperature) input for SWAT hydrologic model.
- Automation codes for satellite images processing (awesome to accomplish multiple tasks with multiple files, I mean 1000s, at a time). For MODIS/AMSRE products use MRT tool or HEG
- Access to gauged data in the US for climatology – NOAA and groundwater and stream flow – USGS
- PRISM precipitation data (nearly 4 km resolution) for the conterminous US
- CHIRPS – up to 0.05 Deg resolution Global rainfall data (1981-present)
- Global 30m resolution DEM from ASTER

Science drivers are often expected to be aligned with broader national research priorities and institutional missions. They reflect the strategic goals and societal challenges that research aims to address. Science drivers also include the need to explore new scientific frontiers, innovate in methodologies, and develop new technologies. This might drive the need for experimental cyberinfrastructure, cutting-edge software development, or novel networking solutions. Modern science often requires collaboration across disciplines, institutions, and even countries. Science drivers in this context might include the need for seamless data sharing, remote access to specialized equipment, or real-time collaboration tools. In addition to research, science drivers may also encompass educational and training needs. This can include the development of educational resources, training programs, and platforms that facilitate learning and skill development in cyberinfrastructure-related fields.

In summary, science drivers in the context of NSF campus cyberinfrastructure programs are the underlying needs and challenges that shape the development of cyberinfrastructure to support scientific research. They are multifaceted and can encompass technological, collaborative, data-related, innovative, educational, and strategic aspects. In this planning project, we have identified more than ten similar science drivers. However, there could be unidentified ones and more importantly, there will be new science drivers. The designed research network has considered scalability and extensibility as important characteristics.

5. Systematic Approach in Designing a Campus Research Network

Designing a research network has its unique requirements and challenges compared to the tasks in enterprise networks. A research network (at least for the case at ISU) aims to provide intra- and inter-campus high-throughput end-to-end network connectivity.

In this planning project, we adopt a well-tested systematic other than ad doc approach called *Analysis-Architecture-Design* [10] to plan and design the research network for ISU. The *ESnet* recommended Science DMZ blueprint [1] has also been used as a guideline for our design.

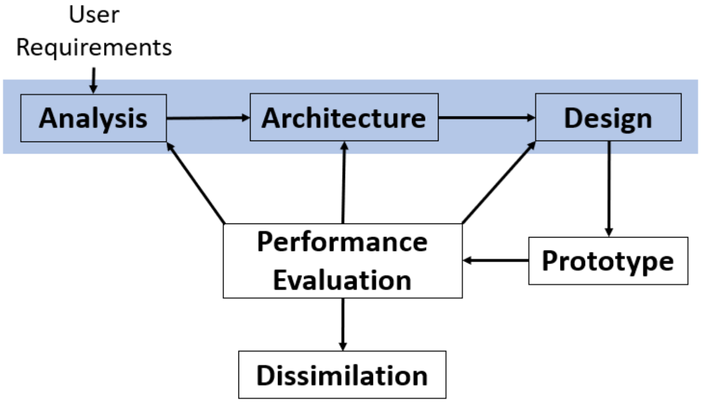

Figure 2: Overview of the Planning Approach

The holistic planning approach is depicted in Fig.2, which consists of the following interactive steps:

- *Analysis*: Gather, derive, define, and validate requirements from science drivers.
- *Architecture*: Determine how addressing and routing, security, network management, and performance are implemented in the network, and how they interact with each other.
- *Design*: Evaluate and select vendors, vendor products, and service providers.
- *Prototype*: Implement a small-scale Science DMZ for one or two selected science drivers.
- *Performance Evaluation*: Evaluate the performance and validate the functionalities of the prototype; evaluation results will be used as new inputs to validate and iteratively conduct, if needed *Analysis-Architecture-Design* and *Prototype* until satisfied.

- *Dissimilation*: The design experience will be dissimilated through meetings and workshops.

Network analysis, architecture, and design are processes used to produce designs that are *logical*, *reproducible,* and *defensible*. These processes are interconnected, in that the output of one process is used directly as input to the next, thus creating flows of information from analysis to architecture, and from architecture to design. Prototyping and performance evaluation have been used to validate and refine the design process iteratively. We are disseminating our design and planning experience to the public. In the following, we elaborate on the details of these interactive steps.

5.1 Who & Why: Requirement Analysis

In this phase, knowing the users (i.e., researchers behind science drivers) and knowing their use cases (i.e., relevant research activities from a high-level conceptual level) are the key, which will guide any technical solution.

In this planning project, we first identified specific science drivers across campus. For each science driver, we adopted Flow Analysis to develop sets of problem statements and objectives that describe what ISU research network should address.

Flows (also known as traffic flows or data flows) are sets of network traffic (application, protocol, and control information) that have common attributes, such as source/destination address, type of information, directionality, or other end-to-end information. Flows are where performance requirements, services, and service metrics are combined with location information to show where performance and service are needed in the network. Flow Analysis provides an end-to-end perspective on requirements and shows where requirements combine and interact. We have examined flows on a link-by-link or network-by-network basis.

For each science driver, we develop a flow specification document called flowspec, which elaborates all relevant flow attributes and performance requirements. This is useful when we want to combine flow requirements from collected flowspecs at the network or network-element levels using different control policies under the Software Defined Networking framework for the research network.

Flow specifications can take one of three types: one-part, or unitary; two-part; or multi-part. Each type of flowspec has a different level of detail, based on whether the flows have best-effort, predictable, and/or guaranteed requirements.

- A one-part flowspec describes flows that have only best-effort requirements.
- A two-part flowspec describes flows that have predictable requirements and may include flows that have best-effort requirements.
- A multi-part flowspec describes flows that have guaranteed requirements and may include flows that have predictable and/or best-effort requirements.

These flow specifications range in complexity. One-part and two-part flowspecs can be relatively straightforward, whereas multi-part flowspecs can be quite complex. Two-part flowspecs are usually a good balance between ease of development and amount of detail. Many networks can be adequately represented with a one-part flowspec, when performance requirements and flows are not well understood. As for the specific purpose research network, however, flows will incorporate more reliability, delay, and throughput requirements, and the two-part and multi-part flowspecs can better represent the research network. This is the case today for many data-intensive research networks. In developing the flowspec we use the information in the analysis of user requirements, and apply the methods described below.

For example, for the research projects conducted in the WRES lab, there contain multiple different types flowspecs across over intra- and inter-campus networks, ranging from the guaranteed requirements needed for their UAS projects, and predictable requirements for their climate sensitivity study, to best-effort for many other projects such as climate impact modeling. For simplicity, we often list projects with different types of flowspecs as different individual science drivers.

Flowspecs are used to combine the performance requirements of multiple science drivers for a composite flow or multiple flows in a section of a path. The flowspec algorithm is a mechanism to combine these performance requirements (capacity, delay, throughput, and reliability) for flows in such a way as to describe the optimal composite performance for that flow or group of flows.

The flowspec algorithm applies the following rules:

- Best-effort flows consist only of capacity requirements; therefore, only capacities are used in best-effort calculations.
- For flows with predictable requirements, we use all available performance requirements (capacity, delay, etc.) in the calculations. Performance requirements are combined for each characteristic so as to maximize the overall performance of each flow.
- For flows with guaranteed requirements, we list each individual requirement (as an individual flow), not combining them with other requirements.

Flow map is also used to describe flows between different end-to-end locations. or network segments to allow better service provisioning and monitoring.

5.2 Research Network Architecture for Measurable Outcomes

Network architecture is the high-level, end-to-end structure of a network. This includes the relationships within and between major architectural components of the network, such as addressing and routing, network management, performance, and security. Determining the network architecture is the next part of the process of developing the campus research network, and is, as we will see, key in integrating requirements and flows into the structure of a network.

A research network is designed to address the limitations of a campus network and is typically deployed near the main campus network. It is important to highlight that the two networks, the research network (aka Science DMZ) and the campus (aka enterprise) network, are separated either physically or logically. There are important reasons for this choice. First, the path from the research network (for both the intra-campus path and the path to the WAN) must involve as few network devices as possible, to minimize the possibility of packet losses at intermediate devices. Second, the research network itself can also be considered a security architecture because it limits the application types and corresponding flows supported by end devices from internal and external networks. While flows in campus networks are numerous and diverse, those in a research network are usually well-identified, enabling security policies to be tied to those flows.

Once the user requirements are clearly elaborated, the next step is to figure out the technology that will help without causing non-productive disruptions. A new mindset even for an IT veteran must be set up is the concept that we are not simply building a network architecture. Instead, we want to build a data architecture. Thus, the project team must throw away a lot of familiar procedures/processes adopted in the traditional network solution space. Designing a research network (i.e., data architecture) implies we must understand the data pipeline for the served researchers, from data creation, data usage, data transferring and sharing, until data curation.

Common high-level research network characteristics are listed below:

- Measurable outcomes: ensure a research network provides expected end-to-end network performance.
- Usable: make the supported researchers can be easily onboarded and integrated into the research network.
- Defensible: control the users and user cases without unnecessarily impacting the usage.
- Scalable: provision of the expected services to current users and keep the space for growing.
- A source of pride: The research network should be a remarkable landmark for an institution to draw more research collaborators and funding.

The main components of a research network are listed below:

- External friction-free network path: DTNs (Data Transfer Nodes, discussed later) are connected to remote systems (e.g., collaborators' networks, data sources or data storage) via the WAN. The high-latency path is composed of routers and switches which have large buffer sizes to absorb transitory packet bursts and prevent losses. The path has no devices that may add excessive delays or cause the packet to be delivered out of order, e.g., firewall, IPS, NAT. The rationale for this design choice is to prevent any packet loss or retransmission which can trigger a decrease in TCP throughput.
- Data Transfer Nodes (DTN): DTNs are dedicated high-performance devices, which are typically Linux devices built and configured for receiving WAN transfers at high speed. They use optimized data transfer tools such as Globus' gridFTP. General-purpose applications (e.g., email clients, MS Office, media players) are not installed. Having a narrow and specific set of applications simplifies the design and enforcement of security policies.
- Performance measurement and monitoring point: Typically, there is a primary high-capacity path connecting the research network with the WAN. An essential aspect is to maintain a healthy path. Identifying and eliminating soft failures in the network is critical for large data transfers. When soft failures occur, basic connectivity continues but high throughput can no longer be achieved. Examples of soft failures include failing components and routers forwarding packets using the main CPU rather than the forwarding plane. Additionally, TCP was intentionally designed to hide transmission errors that may be caused by soft failures. The performance measurement and monitoring point provides an automated mechanism to actively measure end-to-end metrics such as throughput, latency, and packet loss. The most used and widely deployed tool is perfSONAR.
- ACLs (Access Control Lists) and offline security appliances: The primary method to protect a research network is via router's ACLs. Since ACLs are implemented in the forwarding plane of a router or SDN switch, they do not compromise the end-to-end throughput. Additional offline appliances include payload-based and flow-based intrusion detection systems (e.g., open source IDS Zeek).

The architecture uses the information (i.e., flowspecs) from the analysis process to develop a conceptual, high-level, end-to-end structure for the network. Modern network orchestration mechanisms such as

Intent-based networking [41, 42], are being studied to understand their corns and pros in provisioning end-to-end network performance in the planned ISURNet,

In developing the network architecture, we make technology and topology choices for the network. We also determine the relationships among the functions of the network (addressing/routing, network management, performance, and security), and how to optimize the architecture across these relationships. There usually is not a single "right" architecture or design for a network; instead, there are several that will work, some better than others. The architecture and design processes focus on finding those best candidates for architecture and design (optimized across several parameters) for the need of ISU research community.

This ISURNet infrastructure, also termed the Scalable and Polymorphic Research Infrastructure (SPRI), is designed to meet the diverse and dynamic needs of research and education. SPRI comprises three modular components: the Science DMZ (SDMZ) for high-throughput inter-campus networking, the Campus Research Infrastructure (CRI) for a scalable and dynamic intra-campus network, and the Open Programmable Network Platform (OPNP) to foster research innovation. The modular design principle of SPRI ensures that each component meets its unique requirements while enhancing the overall system performance. This initiative is more than just a network development; it's a vision for the future of data-intensive research at ISU.

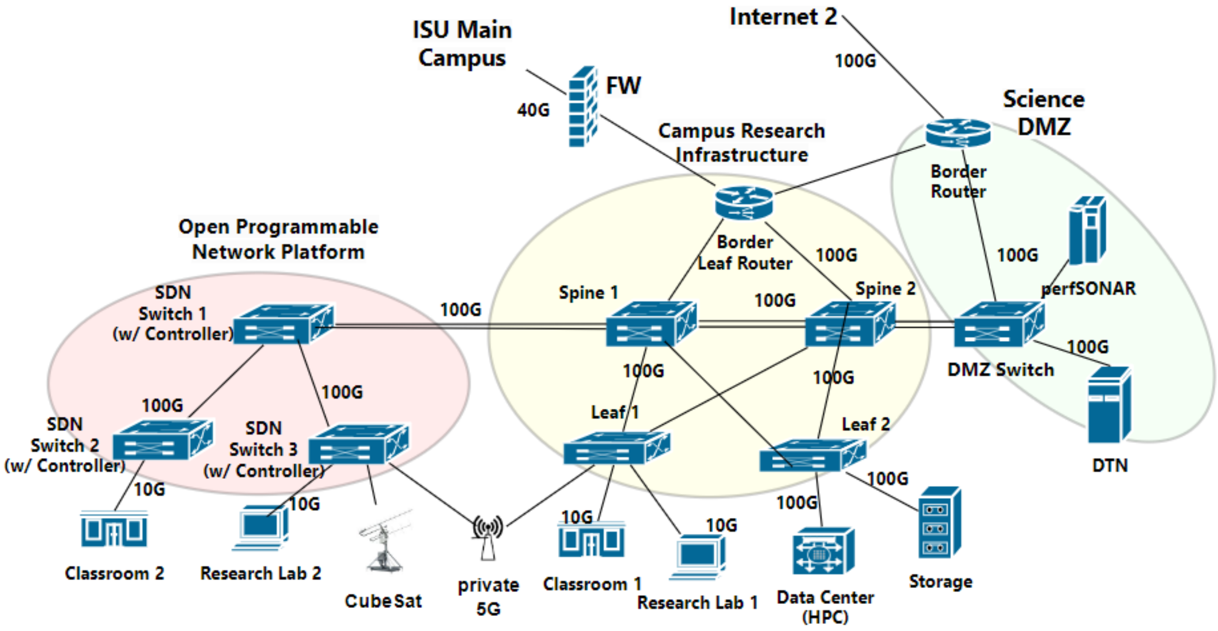

Fig. 3 Proposed Scalable and Polymorphic Research Infrastructure (SPRI)

The goal of SPRI is to tackle the identified issues in the current network infrastructure. SPRI consists of three modular components, and will be fully IPv6 enabled:

SPRI's architecture is carefully crafted, adhering to a modular design principle. This ensures that each component's unique requirements are met while minimizing interference, optimizing overall system performance. Our blueprint is not just about building a network; it's about envisioning the future of data-intensive research and making it a reality.

- Our design for SDMZ is rooted in the ESnet's Science DMZ model, emphasizing frictionless data flow. We've chosen a high-performance switch equipped with a deep buffer, complemented by a top-tier Data Transfer Node (DTN). Additionally, we're integrating a state-of-the-art border router, boasting a dedicated 100 Gbps connection to Internet2 via MREN.
- The CRI is built on the Spine-Leaf architecture, a testament to its scalability, reliability, and high-speed performance. The foundation will consist of two interconnected spine nodes (32x100GbE) using 100Gb links via MLAG, and two leaf nodes. These spines will be strategically positioned in the JH Data Center (Julian Hall) and the STV Campus Data Center (Stevenson Hall). The initial rollout will incorporate 2 Leafs to ensure seamless connectivity to select departments and the HPC infrastructure. Each leaf is designed to support 48x100GbE connections, with dedicated uplinks to the spine switches. We'll leverage existing fiber infrastructure to maximize efficiency. This architecture is primed for swift expansion, ready to support evolving applications and even surpass the speeds of initial endpoints. Furthermore, we're enhancing data transfers to the HPC storage system through an upgraded storage server, which will also serve as the DTN node's storage backbone.
- OPNP stands as a testament to open-ended, adaptable, and scalable design. While it mirrors the Spine-Leaf architecture with a single Spine node currently, its true strength lies in its adaptability. To foster creative research, all nodes in this segment are hybrid SDN switches equipped with built-in controllers. This ensures both centralized and distributed controller architectures can be explored.

This research network aims to dramatically improve data transferring between researchers, scientific instrumentation, visualization workstations, high performance computing (HPC) infrastructure, and external collaborators. Access to this research network enables higher speed to existing research workflows and empowers new research processes previously unavailable to the ISU teams. For example, the research network will facilitate the research collaboration between ISU and the OSF hospital in developing a campus walking assistant to help students with vision problems free walking on campus. Research collaboration between ISU and the OSF hospital applying motion analysis and artificial intelligence to help students with vision problems free walk on campus.

The initial design of ISU research network is a campus-wide 100Gb fiber-based network with Science DMZ, including multiple Data Transfer Nodes (DTNs) enabling data flows separate from the day-to-day traffic of University business. This Architecture process focuses on relationships within each building block and between building blocks, providing an understanding of how each flowspec not only satisfied within a network, but also how it interoperates with other flowspecs. There are numerous trade-offs, dependencies, and constraints that occur between addressing/routing, network management, performance, and security. By manipulating the interactions between flowspecs, the network architecture can be tailored to meet the specific needs as a Science DMZ. The network is implemented as a spine-leaf architecture with two spines interconnected to each other with 100Gb links by multi-chassis link aggregation (MLAG). One spine is placed in the HPC Data Center and the other is placed in a separate campus Data Center. The network has 6 Leafs supporting connectivity between buildings housing academic departments and HPC infrastructure. Performance is monitored and tuned with PerfSonar nodes (not shown in the diagram). Security is monitored through open-source programmable IDS Zeek (40Gb/s) nodes. The research network will be continuously tuned and optimized in collaboration between researchers and network architects.

5.3 Design, Deployment, and Operation

There is constant pressure to deploy new features and services while increasing the quality of existing services and network security in a campus environment. In addition, market forces, and supply-chain risks are pressing network operators to closely manage investment in new infrastructure and decrease operations and maintenance costs. Converting a beautiful network blueprint to a cost-effective and operatable solution is not an easy task anymore.

The design provides physical detail to the architecture. It is the target of our planning project, the culmination of Analysis and Architecture processes. Physical detail includes blueprints and drawings of the network; selections of vendors and service providers, and selections of equipment (including equipment types and configurations).

During network design, we used an evaluation process to make vendor, service provider, and equipment selections, based on input from Analysis and Architecture. We set appropriate design goals, such as minimizing policy intrusiveness and maximizing performance, as well as plan how to achieve these goals, through mapping network performance and function to our design goals and evaluating our design against its goals to recognize when the design varies significantly from these goals. Network design is also about applying the trade-offs, dependencies, and constraints developed as part of the network architecture. Trade-offs, such as security policies versus performance or simplicity versus function, occur throughout the design process, and a large part of network design concerns recognizing such trade-offs (as well as interactions, dependencies, and constraints) and optimizing the design among them. As part of the design process, we will also learn how to develop evaluation criteria for our design.

Several vendors have been carefully evaluated. For example, 7280R3 product line from Arista Networks can provide full-feature routing and SDN functionalities with 100/400G wire speed and deep buffers, which could be candidate equipment for the ISU research network. Its built-in network telemetry functions provide a complementary monitoring solution for intra-campus network segments, along with perfSONAR for inter-campus network segments. Furthermore, its multi-function NetOps platform called CloudVision can provide a suitable orchestration among all devices in the research network, including automated deployments, real-time telemetry, change controls, and security services.

5.4 Prototype

The ESnet Science DMZ model [1] allows new services to be tested, validated, and then rolled into production once they are proven operationally sound. In this planning project, we have prototyped a small-scale designed Science DMZ for one simulated science driver. The purpose of prototyping is to test related data transfer tools and services (e.g., Globus [9]), and validate the flexibility, extensibility, performance, and incremental deployment strategy of the designed research network. More specifically, the prototyping simply uses two Linux servers connected via an Arista 7280TR-48C6 SDN switch. We used the network emulation tool (called netem) to simulate WAN (I2): delay, jitter, and packet losses to 1) understand how to set up buffers on DTN and switch/router; 2) understand the relation between buffer size and bandwidth-delay product (BDP); 3) test the functionalities of Globus & perfSonar under different network conditions (e.g., congestion); 4) understand how the firewall rules affect network throughput.

The prototyping is still under investigation now. The performance test results from the prototyped system could facilitate the equipment selection and configuration.

6. The Impact

This project facilitated various interdisciplinary collaborations, uniting institutional administrators, IT practitioners, and academic researchers in a joint exploration of the unique interplay between scientific and educational imperatives and their cyberinfrastructure needs. The project engaged participants recognized designing a research network for Primarily Undergraduate Institutions (PUIs) like ISU presents unique challenges and considerations compared to larger research universities, including but not limited to limited resources, a diversity of research agendas, a strong educational focus, and a deep connection with local communities.

We conducted a detailed analysis of ISU's science drivers, offering insights into the specific needs and opportunities at such institutions. The outcome was the creation of a Scalable and Polymorphic Cyberinfrastructure – a dedicated but flexible research network optimized for high-performance and innovative scientific applications and data transfer. The relevant findings and experiences have been disseminated via multiple workshops and a publicly accessible website to IT staff and researchers in higher education, especially those from PUIs.

Aligning CI planning with the wide breadth of the research and education community highlights a broader impact on ISU's strategic plan. This project plans and promises a compelling cyberinfrastructure (CI) for the STEM community at ISU. The two-way conversations in the planning phase between researchers and CI professionals are not an easy but essential step for successful and meaningful broader research results.

Moreover, the increased awareness will be emitted from ISU as a hub via various dissemination methods (e.g., workshops, websites) to the local community, the region, and many connected institutions nationally, such as the member institutions in Intercollegiate Biomathematics Alliance led by ISU. its success will not only position ISU as a frontrunner in undergraduate education but also stand as a model CI blueprint, tailored for PUIs to meet their diverse research and educational demands in data-intensive computing. It offers a viable solution for PUIs, often an overlooked segment in the US education system, by streamlining the CI ecosystem. Moreover, it aims to enrich the academic journey of underrepresented students across various scientific domains.

References

[1] E. Dart, L. Rotman, B. Tierney, M. Hester and J. Zurawski, "The Science DMZ: A network design pattern for data-intensive science," SC '13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, 2013, pp. 1-10, doi: 10.1145/2503210.2503245.

[2] Illinois State University Strategic Plan 2018-2023. https://strategicplan.illinoisstate.edu/ Last accessed on 8/10/2023.

[3] FABRIC National Research Infrastructure. https://whatisfabric.net/ Last accessed on 8/10/2023.

[4] The Engagement and Performance Operations Center (EPOC). https://epoc.global/cc/ Last accessed on 8/10/2023.

[5] Energy Sciences Network (ESnet): https://fasterdata.es.net/science-dmz Last accessed on 8/10/2023.

[6] Chameleon: A Configurable Experimental Environment for Large-scale Edge to Cloud Research. https://www.chameleoncloud.org/ Last accessed on 8/10/2023.

[7] CloudLab: A Flexible, scientific infrastructure for research on the future of cloud computing. https://cloudlab.us/ Last accessed on 8/10/2023.

[8] Jetstream: A User-friendly Cloud Computing Environment for Researchers. https://jetstream-cloud.org/ Last accessed on 8/10/2023.

[9] Globus: A Data-intensive Transfer Tool. https://www.globus.org/ Last accessed on 8/10/2023.

[10] James D. McCabe. "Network Analysis, Architecture, and Design". Morgan Kaufmann, 3rd Edition, 2007.

[11] The Extreme Science and Engineering Discovery Environment (XSEDE). https://www.xsede.org/ Last accessed on 8/10/2023.

[12] Open Science Grid (OSG). https://opensciencegrid.org/ Last accessed on 8/10/2023.

[13] Pacific Research Platform (PRP). https://prp.lsmarr.net/ Last accessed on 8/10/2023.

[14] Global Environment for Network Innovations (GENI) https://www.geni.net/ Last accessed on 8/10/2023.

[15] Junhong Xu, Shangyue Zhu, Hanqing Guo, Shaoen Wu. "Automated Labeling for Robotic Autonomous Navigation Through Multi-Sensory Semi-Supervised Learning on Big Data". IEEE Transactions on Big Data. 2019.

[16] Hanqing Guo, Junhong Xu, Shangyue Zhu, Shaoen Wu." Realtime Software Defined Self-Interference Cancellation Based on Machine Learning for In-Band Full Duplex Wireless Communications". International Conference on Computing, Networking and Communications (ICNC). 2018.

[17] Hanqing Guo, Nan Zhang, Shaoen Wu, Qing Yang. "Deep Learning Driven Wireless Real-time Human Activity Recognition". IEEE International Conference on Communications (ICC). 2020.

[18] Erwin Cornelius, Olcay Akman, Dan Hrozencik. "COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty". Journal of Mathematics. 2021.

[19] Anuj Mubayi, Jeff Sullivan, Jason Shafrin, Oliver Diaz, Aditi Ghosh, Anamika Mubayi, Olcay Akman, Phani Veeranki. "Battling Epidemics & Disparity with Modeling". Letters in Biomathematics. 2020.

[20] Jorge Humberto Rojas, Marlio Paredes, Malay Banerjee, Olcay Akman, Anuj Mubayi. "Mathematical modeling & the transmission dynamics of SARS-CoV-2 in Cali, Colombia: implications to a 2020 outbreak & public health preparedness". Journal of medRxiv. 2020.

[21] Y. Lu, N. Chris- tensen, Q. Su and R. Grobe, "Space-time resolved Breit-Wheeler process for a model system," Physics Review A101, 022503 (2020).

[22] N. Christensen, J. Unger, S. Pinto, Q. Su and R. Grobe, "Spatial Evolution of Quantum Mechanical States," Annals of Physics 389 (2018) 239-249.

[23] McGinnis, C., Holland, D., Su, Q., & Grobe, R. Universal scaling laws for optimally excited nonlinear oscillators. Phys Rev E (2020)

[24] Su, Q. Resolving rapidly chirped external fields with Dirac vacuum. Phys. Rev. A 101 (2020): 063405.

[25] NSF RUI: Spatial and Temporal Dynamics of Matter in Intense Laser Fields https://www.nsf.gov/awardsearch/showAward?AWD_ID=2106585&HistoricalAwards=false

[26] Abdelmounaam Rezgui, Nickolas Davis, Zaki Malik, Brahim Medjahed, Hamdy S Soliman. "CloudFinder: A system for processing big data workloads on volunteered federated clouds". IEEE Transactions on Big Data. 2017.

[27] Nickolas Allen Davis, Abdelmounaam Rezgui, Hamdy Soliman, Skyler Manzanares, Milagre Coates. "Failuresim: A system for predicting hardware failures in cloud data centers using neural networks". IEEE 10th International Conference on Cloud Computing (CLOUD). 2017.

[28] Abdelmounaam Rezgui, Kyoomars Alizadeh Noghani, Javid Taheri. "SDN helps big data to become fault tolerant". IET Digital Library. 2018.

[29] Amir Mirzaeinia, Abdelmounaam Rezgui, Zaki Malik, Mehdi Mirzaeinia. "Min-edge p-cycles: an efficient approach for computing p-cycles in optical data center networks". nternational Conference on Cloud Computing. 2019.

[30] Muteb Alshammari, Abdelmounaam Rezgui. "POX-PLUS: An SDN Controller with Dynamic Shortest Path Routing". IEEE 9th International Conference on Cloud Networking (CloudNet). 2020.

[31] Guang Cheng and Yongning Tang, "eOpenFlow: Software Defined Sampling via a Highly Adoptable OpenFlow Extension". IEEE International Conference on Communications. Paris, France, May 21-25, 2017

[32] Yongning Tang, Guang Cheng, Zhiwei Xu, Feng Chen, Khalid Elmansor, and Yangxuan Wu, "Automatic Belief Network Modelling via Policy Inference for SDN Fault Localization". (2016) Journal of Internet Services and Applications. Vol. 7:1. DOI: 10.1186/s13174-016-0043-y

[33] Yongning Tang, Yangxuan Wu, Guang Cheng, Zhiwei Xu. "ntelligence enabled sdn fault localization via programmable in-band network telemetry". IEEE 20th International Conference on High Performance Switching and Routing (HPSR). 2019.

[34] Guang Cheng, Chunsheng Guo, Yongning Tang. "dptCry: an approach to decrypting ransomware WannaCry based on API hooking". CCF Transactions on Networking. 2019.

[35] Ying Hu, Guang Cheng, Yongning Tang, Feng Wang. "A practical design of hash functions for IPv6 using multi-objective genetic programming". Computer Communications. 2020.

[36] Feng Wang, Yongning Tang, Lixin Gao, Guang Cheng. "BC-Sketch: A Simple Reversible Sketch for Detecting Network Anomalies". IEEE International Conference on Smart Data Services (SMDS). 2020.

[37] Feng Wang, Eduard Babulak, Yongning Tang, "On Detecting and Identifying Faulty IoT Devices and Outage". Bulletin of Electrical Engineering and Informatics, 2021.

[38] Jessil Fuhr, Isaac Hanna, Feng Wang, Yongning Tang, "The Diminished Importance of Connection-based Features in Intrusion Detection". IEEE International Performance, Computing, and Communications Conference, 2021.

[39] A. Siddiqui, J. Qaddour and S. Ullah, "Securing Healthcare IoT (HIoT) Monitoring System Using Blockchain," 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), 2021, pp. 60-66, doi: 10.1109/ICUFN49451.2021.9528625.

[40] Amirata Ghorbani, James Zou. "Data Shapley: Equitable Valuation of Data for Machine Learning". https://arxiv.org/pdf/1904.02868.pdf Last accessed on 10/10/2021.

[41] S. Alalmaei, Y. Elkhatib, M. Bezahaf, M. Broadbent and N. Race, "SDN Heading North: Towards a Declarative Intent-based Northbound Interface," 2020 16th International Conference on Network and Service Management (CNSM), 2020, pp. 1-5, doi: 10.23919/CNSM50824.2020.9269118.

[42] N. Herbaut, C. Correa, J. Robin and R. Mazo, "SDN Intent-based conformance checking: application to security policies," 2021 IEEE 7th International Conference on Network Softwarization (NetSoft), 2021, pp. 181-185, doi: 10.1109/NetSoft51509.2021.9492679.

[43] NSF CC Planning: Fostering Data Driven STEM Research and Education Through Cyberinfrastructure at Illinois State University https://www.nsf.gov/awardsearch/showAward?AWD_ID=2201478&HistoricalAwards=false*

[44] Yost, J., Rizo, L., Fang, X., Su, Q., & Grobe, R. (2022). Exactly predictable functions for simple neural networks. SN Computer Science, 3, 1-13.

[45] Gong, C., Bryan, J., Furcoiu, A., Su, Q., & Grobe, R. (2022). Evolutionary symbolic regression from a probabilistic perspective. SN Computer Science, 3(3), 209.

[46] Lv, Q. Z., Jennings, D. J., Betke, J., Su, Q., & Grobe, R. (2016). Optimized spatial matrix representations of quantum Hamiltonians. Computer Physics Communications, 198, 31-40.

[47] Norris, S., Vikartofsky, A., Wagner, R. E., Su, Q., & Grobe, R. (2013). Absorbing-like boundaries for quantum field theoretical grid simulations. Computer Physics Communications, 184(11), 2412-2418.

[48] Xuemei Han, Aaron Aslanian, and John R Yates. Mass spectrometry for proteomics. Current Opinion in Chemical Biology, 12(5):483–490, 10 2008.

[49] Ajith Kumar, Vaishali Narayanan, and Ashok Sekhar. Characterizing post-translational modifications and their effects on protein conformation using nmr spectroscopy. Biochemistry, 59:57–73, 2020. publisher: American Chemical Society Citation Key: Kumar2020.

[50] Kenneth Lucas and George L. Barnes. Modeling the effects of o-sulfonation on the cid of serine. Journal of the American Society for Mass Spectrometry, 31(5):1114–1122, 5 2020. PMID: 32202776 publisher: NLM (Medline) Citation Key: Lucas2020.

[51] Kenneth Lucas, Amy Chen, Megan Schubmehl, Kristopher J. Kolonko, and George L. Barnes. Exploring the effects of methylation on the cid of protonated lysine: A combined experimental and computational approach. Journal of the American Society for Mass Spectrometry, 32:2675–2684, 10 2021. publisher: American Chemical Society (ACS) Citation Key: Lucas2021.

[52] Fred W. McLafferty, Kathrin Breuker, Mi Jin, Xuemei Han, Giuseppe Infusini, Honghai Jiang, Xianglei Kong, and Tadhg P. Begley. Top-down ms, a powerful complement to the high capabilities of proteolysis proteomics. The FEBS Journal, 274(24):6256–6268, 2007. e print : htt ps : //onlinelibrary.wiley.com/doi/pd f /10.1111/ j.1742−4658.2007.06147.x.

[53] Ole Norregaard Jensen. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Current Opinion in Chemical Biology, 8(1):33–41, 2 2004.

[54] B´ela Paizs and S´andor Suhai. Fragmentation pathways of protonated peptides. Mass spectrometry reviews, 24(4):508–548, 2005. PMID: 15389847 Citation Key: Paizs2005.

[55] Subha Pratihar, George L. Barnes, William L. Hase, Subha Pratihara, George L. Barnes, and William L. Hase. Chemical dynamics simulations of energy transfer, surface-induced dissociation, soft-landing, and reactive- landing in collisions of protonated peptide ions with organic surfaces. Chemical Society reviews, 45(13):3595–3608, 11 2016. publisher: The Royal Society of Chemistry Citation Key: Pratihar2016a.

[56] Subha Pratihar, George L. Barnes, Julia Laskin, and William Louis Hase. Dynamics of protonated peptide ion collisions with organic surfaces. consonance of simulation and experiment. The journal of physical chemistry letters, 7:3142–3150, 8 2016. PMID: 27467857 publisher: American Chemical Society Citation Key: Pratihar2016b.

[57] Subha Pratihar, Xinyou Ma, Zahra Homayoon, George L. Barnes, andWilliam L. Hase. Direct chemical dynamics simulations. Journal of the American Chemical Society, 139(10):3570–3590, 3 2017. publisher: American Chemical Society

[58] Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. Database, 2021:baab012, 9 2021. Citation Key: Ramazi2021

[59] Mikhail M. Savitski, Maria F¨alth, Y. M. Eva Fung, Christopher M. Adams, and Roman A. Zubarev. Bifurcating fragmentation behavior of gas-phase tryptic peptide dications in collisional activation. Journal of the American Society for Mass Spectrometry, 19(12):1755–1763, 12 2008. publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[60] Jun Feng Xiao, Bin Zhou, and Habtom W. Ressom. Metabolite identification and quantitation in lc-ms/ms-based metabolomics. TrAC Trends in Analytical Chemistry, 32:1–14, 2 2012.

[61] Talat Yalcin, Imre G. Csizmadia, Michael R. Peterson, and Alex G. Harrison. The structure and fragmentation of bn (n3) ions in peptide spectra. Journal of the American Society for Mass Spectrometry, 7(3):233–242, 3 1996. publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[62] Talat Yalcin, Charlotte Khouw, Imre G. Csizmadia, Michael R. Peterson, and Alex G. Harrison. Why are b ions stable species in peptide spectra? Journal of the American Society for Mass Spectrometry, 6(12):1165–1174, 12 1995. publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved

[63] Kristen L. Reese, A. Daniel Jones, and Ruth Waddell Smith. Characterization of smokeless powders using multiplexed collision-induced dissociation mass spectrometry and chemometric procedures. Forensic Science International, 272:16–27, 3 2017.