I am a *computer systems researcher focused on improving the sustainability of computing*. My work pushes the boundaries of computer systems design and operation to address emerging challenges of rapidly rising computing demand, increasing energy availability constraints, and unintended socio-environmental implications of computing. These are challenges our current infrastructure cannot solve. I take the requisite multidisciplinary approach that integrates domain-specific knowledge from energy systems and industrial ecology with advanced computer systems approaches to develop high-impact solutions at all layers of computer system stacks and all steps in their lifecycles. In manifesting real-world impact, my work has enhanced the resource efficiency of hyperscale datacenters [31] and powered community testbeds for carbon-efficient applications [47].

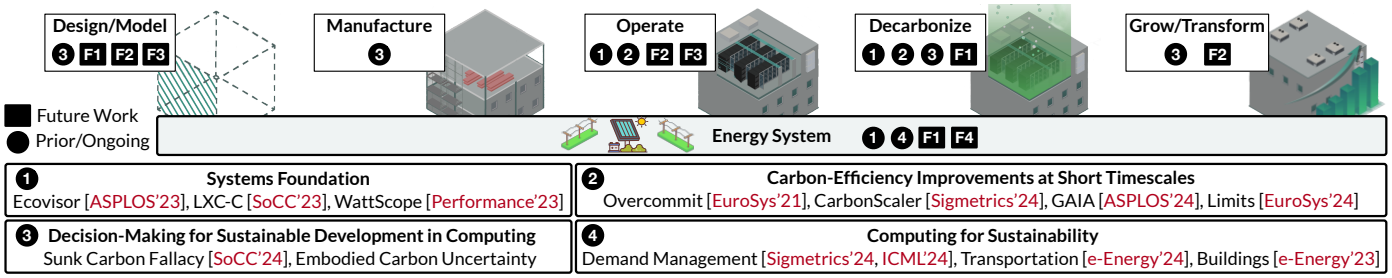

Fig.1: **An overview of my prior (❶ – ❹) and future work (F1 – F4) across computer systems' lifecycle stages.**

Research Overview Over the past two decades, energy efficiency optimizations have increased computing's economic productivity but have not reduced its aggregate energy demand or environmental impact. Moreover, we are approaching several physical limits of energy efficiency that have tempered growth in energy demand, with the end of Dennard Scaling and the slowdown of Moore's Law [48, 49]. Moving forward, *mitigating computing's lifecycle environmental impact will require prioritizing carbon efficiency – measured by the work done per unit of carbon (and other greenhouse gases) emitted.* Computing's operational emissions (produced by energy use) can be reduced by doing more work when and where low-carbon energy is available. Reducing embodied carbon emissions (from the production and disposal of computing hardware and infrastructure) necessitates reevaluating hardware design, procurement, and capacity provisioning strategies. Ultimately, improving carbon efficiency requires fundamentally new and disruptive research across the computer system's software and hardware stack, including modeling, design, manufacturing, operation, and graceful lifecycle extension of computing infrastructure. Figure 1 shows the contributions I have made towards solving these challenges. I briefly overview representative contributions below.

❶ **Systems Foundations.** Computing applications lack visibility and control over their energy supply, preventing them from adjusting power usage based on energy's carbon intensity or renewable energy availability. To address this issue, I have *done foundational work on virtualizing datacenter energy systems and exposing software-defined control to applications* [26]. Inspired by Exokernel, these abstractions enable applications to manage clean energy's variability within their software stack directly, aligning performance needs with sustainability goals by leveraging one or more dimensions of software flexibility and fault tolerance. `Ecovisor`'s software ecosystem is open-source and deployed on a community testbed [27, 42, 45, 47]. I have also developed supporting tools to enable non-intrusive energy monitoring, thermal energy management, and a fair distributed rate control for the Ecovisor ecosystem [18, 19, 39].

❷ **Carbon Efficiency Improvements at Short Timescales.** At seconds to days timescales, improving carbon efficiency requires continuously optimizing workload execution. My work highlighted that simultaneously optimizing for carbon, energy, and performance is impossible. Using this insight, I have *designed systems for various applications that strategically trade energy or performance to achieve carbon-efficiency improvements.* For instance, I developed `CarbonScaler` [20], a carbon-aware autoscaler that scales up (energy inefficiently) during low-carbon periods (carbon efficient) without increasing job completion time (maintaining performance). `CarbonScaler` and related artifacts are open-source [43, 45]. I demonstrated that carbon footprint reduction opportunities are available for a broad set of workloads – such as data processing, scientific computing, and AI training – and in various computing environments, such as public clouds, on-premise datacenters, edge computing, and hybrid clouds [3, 6, 11, 20, 23, 24, 30, 34].

❸ **Decision-Making for Sustainable Development in Computing.** Improving computing's carbon efficiency through sustainable choices across lifecycle stages – chip design, server procurement, and datacenter siting – often relies on data with significant uncertainties. I have *quantified uncertainty in carbon estimates, shown its impact on decision discernibility, and developed strategies for decision-making under uncertainty.* In doing so, I extended the Product Attributes and Impact Algorithm (PAIA), a lifecycle analysis tool for ICT companies [44], to support uncertainty-driven quantitative assessments of decisions [4]. Furthermore, the metrics quantifying computing's carbon footprint and making decisions are still evolving. I have also rigorously evaluated carbon-based metrics and their incentives for holistic carbon reduction [1, 2, 5, 10, 16, 28].

❹ **Computing for Sustainability.** The net sustainability implications of computing and AI tools can be improved by using them to quantify and reduce emissions in other societal sectors. I have done extensive work on using computational methods to accelerate the decarbonization of the electric grid, buildings, and transportation sectors. First, I have *designed spatiotemporal scheduling algorithms for learning-augmented online optimization that use AI predictions to achieve better average-case performance without sacrificing worst-case competitive guarantees* [7–9, 22]. These algorithms apply to a broad set of sustainability problems, such as carbon-aware electric vehicle charging and computing workload execution. Second, I *used physical models and data-driven ML methods* for solar PV performance modeling and forecasting [32, 35, 37], Bayesian methods for anomaly detection in solar panels [33, 36], and distributed rate control approaches from computer networks for controlling distributed solar capacity [38, 39]. Third, to support energy transition in the buildings sector, I have developed tools for *tactical energy transition* from gas-based heating to electric heat pumps [21, 29], incentive design for solar energy adoption [14, 15, 25], and devising smart load-shedding solutions [40, 41]. Finally, I have devised *carbon- and equity-aware ride assignment policies* for ridesharing platforms [12, 13, 17].

Across these threads, I have enabled computing stakeholders to reliably quantify and significantly reduce the lifecycle environmental impact of AI demand while demonstrating how AI can accelerate societal decarbonization. In doing so, I utilized the system stack of software-defined infrastructure, distributed systems, resource management, and performance evaluation.

Future Research Directions My work on building systems, carbon-efficient applications, and frameworks for sustainable AI is evergreen. However, its capabilities must expand, and its application to improving computing's carbon efficiency will need to evolve, presenting a rich set of challenges as the insatiable demand for AI workloads grows, new application frameworks emerge, tangible incentives to reduce carbon footprints are introduced, or planetary limits on materials and emissions are reached. Below, I envision my research addressing existing and future challenges in enabling sustainable AI, shown in Figure 1.

F1 – Designing and Operating Sustainable Datacenters. Meeting the growing demand for AI workloads responsibly requires a sustainability-aware, multidisciplinary approach that both looks inward to address user, application, and infrastructure challenges in datacenters and outward to understand electric grid constraints and the challenges it faces for a reliable operation.

My work will lay the groundwork for sustainable datacenters by benchmarking their architecture, hardware, and workloads to assess tradeoffs between performance, energy, and sustainability metrics, such as carbon and water footprint. In this effort, I will develop system support that enables datacenters to automatically adjust their operations at minimum performance impact based on grid conditions at short time scales while optimizing design for mutually beneficial coordination with the electric grid over the long term. I will develop higher-level frameworks for the sustainable operation of datacenters, respecting the constraints and objectives of users, datacenter operators, and grid utilities. This will involve creating carbon-centered service level agreements (cSLAs) that allow cloud platforms to offer sustainable solutions while enabling users to optimize sustainability and financial goals. Finally, I will develop frameworks for datacenter–grid coordination, leveraging game-theoretic approaches to design datacenter demand response solutions that offer meaningful incentives for participation.

F2 – Co-Adapting Emerging Applications and Heterogeneous Hardware. Modern cloud-native and AI inference-driven applications are being deployed on heterogeneous (specialized and aging) hardware and are driving much of the increase in computing demand. Prior work does not tackle optimizing the carbon efficiency of this evolving software-hardware ecosystem.

Interestingly, the defining aspects of emerging applications – scalability, resiliency, and redundancy – are also desired characteristics for using power from highly variable and unreliable renewable energy sources. I will develop new abstractions for developers and cloud operators to deploy modern applications on specialized (for performance) and aging (to reduce embodied carbon) hardware run on intermittent renewable power (to reduce operational carbon) while balancing ease of use against deep optimizations. However, these applications' sheer scale and distributed nature make deploying any optimizations challenging, and simple heuristics do not work well. While modern black-box ML/AI tools for systems are being used, they are generally reserved for non-critical and particular use cases due to their poor generalization and lack of worst-case guarantees. I will continue my work on combining AI advice with robust algorithms to provide good average-case performance and worst-case guarantees when using AI tools for systems, ensuring they remain adaptive and robust in dynamic and uncertain environments.

F3 – Digital Twins-in-the-Loop (DTIL) Datacenters. There is no production-scale deployment of sustainable computing solutions, such as spatiotemporal workload migrations and datacenter demand response. Beyond the lack of incentives, the key obstacle is an obscured view of sustainability-driven optimizations' financial, technical, and infrastructure implications.

My research will lead an effort to build holistic end-to-end models for distributed datacenter infrastructure that accounts for cost, energy, carbon, and performance impacts of carbon-aware optimizations. For instance, I will develop realistic models for workload migrations' cost and carbon impacts in the network. Similarly, I will also focus on the challenging and pressing issue of modeling the tradeoffs between redundancy, availability, and embodied carbon. The modeling effort will drive a reassessment of the relentless pursuit of marginal gains in performance – without demonstrated system-scale benefits – and identify the

realistic carbon-aware optimizations at scale. In the long run, guided by the modeling efforts, I aim to leverage AI tools to design digital twin-in-the-loop (DTiL) datacenters that have a symbiotic relationship with the physical infrastructure. The DTiL datacenters will use digital twins to optimize operations at short timescales and inform design at long timescales.

F4 – Computing-Energy-Society Nexus. A push towards electrification and embedded intelligence across various societal sectors creates interdependencies that did not exist before. For instance, smart electric cars are changing the landscape of personal transportation: they require computing resources across the stack (device, edge, cloud) and create electricity couplings between residential and commercial buildings. Ultimately, datacenters, the electric grid, and other societal sectors are increasingly coupled due to reliance on computing and electric grids. This means that a siloed focus on improving resource efficiency in each sector is unlikely to yield practical solutions that lead to societal-scale decarbonization [46]. My research will design computing solutions (infrastructure and software) that power computational approaches to holistic cross-domain decarbonization. The geographical distribution of demand (e.g., roadside, buildings, cloud-based) and its varying temporal characteristics (e.g., ephemeral on the roadside, periodic in buildings, continuous analytics in the cloud) will require hardware-software solutions that sustainably serve cross-sectoral approaches instead of domain-specific over-provisioned solutions.

References

[1] **Noman Bashir**, Priya Donti, James Cuff, Sydney Sroka, Marija Ilic, Vivienne Sze, Christina Delimitrou, and Elsa Olivetti. "The Climate and Sustainability Implications of Generative AI". In: An MIT Exploration of Generative AI. MIT Press, 2024.

[2] **Noman Bashir**, Varun Gohil, Mohammad Shahrad, David Irwin, Anagha B. Subramanya, Elsa Olivetti, and Christina Delimitrou. "The Sunk Carbon Fallacy: Rethinking Carbon Footprint Metrics for Effective Carbon-Aware Scheduling". In: ACM SoCC. 2024.

[3] **Noman Bashir**, Adam Lechowicz, Rohan Shenoy, Mohammad Hajiesmaili, Adam Wierman, and Christina Delimitrou. "Learning Carbon-Aware Scheduling Algorithms for Data Processing Clusters". In: 2024.

[4] **Noman Bashir**, Anagha B. Subramanya, Julia Xia, Varun Gohil, Ajay Gupta, Melissa Zgola, Greg Norris, Elsa Olivetti, and Christina Delimitrou. "Discernible Decision Making under Uncertainty in Sustainable Computing". In: 2024.

[5] Yichen Gao, **Noman Bashir**, Christopher Hill, and Jeremy Gregory. "Enabling Proactive Sustainability Interventions in Datacenters". In: submission. 2024.

[6] Walid Hanafy, Qianlin Liang, **Noman Bashir**, Abel Souza, David Irwin, and Prashant Shenoy. "Going Green for Less Green: Optimizing the Cost of Reducing Cloud Carbon Emissions". In: ACM ASPLOS. 2024.

[7] Adam Lechowicz, Nicolas Christianson, Bo Sun, **Noman Bashir**, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. "CarbonClipper: Optimal Algorithms for Carbon-aware Spatiotemporal Workload Management". In: submission. 2024.

[8] Adam Lechowicz, Nicolas Christianson, Bo Sun, **Noman Bashir**, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. "Online Conversion with Switching Costs: Robust and Learning-Augmented Algorithms". In: ACM SIGMETRICS. 2024.

[9] Adam Lechowicz, Nicolas Christianson, Bo Sun, **Noman Bashir**, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. "Chasing Convex Functions with Long-term Constraints". In: ICML. 2024.

[10] Diptyaroop Maji, **Noman Bashir**, David Irwin, Prashant Shenoy, and Ramesh K Sitaraman. "The Green Mirage: Impact of Location- and Market-based Carbon Intensity Estimation on Carbon Optimization Efficacy". In: ACM e-Energy. 2024.

[11] Talha Mehboob, **Noman Bashir**, Jesus Omana Iglesias, Michael Zink, and David Irwin. "EcoLearn: Optimizing the Carbon Footprint of Federated Learning". In: submission. 2024.

[12] Mahsa Sahebdel, Ali Zeynali, **Noman Bashir**, Prashant Shenoy, and Mohammad Hajiesmaili. "LEAD: Towards Learning-Based Equity-Aware Decarbonization in Ridesharing Platforms". In: submission. 2024.

[13] Mahsa Sahebdel, Ali Zeynali, **Noman Bashir**, Prashant Shenoy, and Mohammad Hajiesmaili. "A Holistic Approach for Equity-aware Carbon Reduction of the Ridesharing Platforms". In: ACM e-Energy. 2024.

[14] Cooper Sigrist, Adam Lechowicz, Jovan Champ, **Noman Bashir**, and Mohammad Hajiesmaili. "Lost in Siting: The Hidden Carbon Cost of Inequitable Residential Solar Installations". In: submission. 2024.

[15] Anupama Sitaraman, Adam Lechowicz, **Noman Bashir**, Xutong Liu, Mohammad Hajiesmaili, and Prashant Shenoy. "Dynamic Incentive Allocation for City-Scale Deep Decarbonization". In: submission. 2024.

[16] Thanathorn Sukprasert, **Noman Bashir**, Abel Souza, David Irwin, and Prashant Shenoy. "On the Implications of Choosing Average versus Marginal Carbon Intensity Signals on Carbon-aware Optimizations". In: ACM e-Energy. 2024.

[17] Ali Zeynali, Mahsa Sahebdel, **Noman Bashir**, Ramesh Sitaraman, and Mohammad Hajiesmaili. "Near-Optimal Emission-Aware Online Ride Assignment Algorithm for Peak Demand Hours". In: submission. 2024.

[18] **Noman Bashir**, Yasra Chandio, David Irwin, Fatima M. Anwar, Jeremy Gummeson, and Prashant Shenoy. "Jointly Managing Electrical and Thermal Energy in Solar- and Battery-powered Computer Systems". In: ACM e-Energy. 2023.

[19] Xiaoding Guan, **Noman Bashir**, David Irwin, and Prashant Shenoy. "WattScope: Non-intrusive Application-level Power Disaggregation in Datacenters". In: IFIP Performance. 2023.

[20] Walid Hanafy, Qianlin Liang, **Noman Bashir**, David Irwin, and Prashant Shenoy. "CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing Carbon-Efficiency". In: ACM SIGMETRICS. 2023. **Best Student Paper Award**.

[21] Adam Lechowicz, **Noman Bashir**, John Wamburu, Mohammad Hajiesmaili, and Prashant Shenoy. "Equitable Network-Aware Decarbonization of Residential Heating at City Scale". In: ACM e-Energy. 2023.

[22] Adam Lechowicz, Nicolas Christianson, Jinhang Zuo, **Noman Bashir**, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. "The Online Pause and Resume Problem: Optimal Algorithms and An Application to Carbon-Aware Load Shifting". In: ACM SIGMETRICS. 2023.

[23] Qianlin Liang, Walid Hanafy, **Noman Bashir**, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. "Dělen: Enabling Flexible and Adaptive Model-serving for Multi-tenant Edge AI". In: ACM/IEEE IoTDI. 2023.

[24] Qianlin Liang, Walid Hanafy, **Noman Bashir**, David Irwin, and Prashant Shenoy. "Energy Time Fairness: Balancing Fair Allocation of Energy and Time for GPU Workloads". In: IEEE/ACM SEC. 2023.

[25] Anupama Sitaraman, **Noman Bashir**, David Irwin, and Prashant Shenoy. "No Free Lunch: Analyzing the Cost of Deep Decarbonizing Residential Heating Systems". In: IGSC. 2023. **Best Student Paper Award**.

[26] Abel Souza, **Noman Bashir**, Jorge Murillo, Walid Hanafy, Qianlin Liang, David Irwin, and Prashant Shenoy. "Ecovisor: A Virtual Energy System for Carbon-Efficient Applications". In: ACM ASPLOS. 2023.

[27] John Thiede, **Noman Bashir**, David Irwin, and Prashant Shenoy. "Carbon Containers: A System-level Facility for Managing Application-level Carbon Emissions". In: ACM SoCC. 2023.

[28] **Noman Bashir**, David Irwin, Prashant Shenoy, and Abel Souza. "Sustainable Computing – Without the Hot Air". In: HotCarbon. 2022.

[29] John Wamburu, **Noman Bashir**, David Irwin, and Prashant Shenoy. "Data-driven Decarbonization of Residential Heating Systems". In: ACM BuildSys. 2022.

[30] Pradeep Ambati, **Noman Bashir**, David Irwin, and Prashant Shenoy. "Good Things Come to Those Who Wait: Optimizing Job Waiting in the Cloud". In: ACM SoCC. 2021.

[31] **Noman Bashir**, Nan Deng, Krzysztof Rzadca, David Irwin, Sree Kodak, and Rohit Jnagal. "Take it to the Limit: Peak Prediction-driven Resource Overcommitment in Datacenters". In: ACM EuroSys. 2021.

[32] **Noman Bashir**, David Irwin, and Prashant Shenoy. "A Probabilistic Approach to Committing Solar Energy in Day-ahead Electricity Markets". In: IGSC/SUSCOM (2021).

[33] Menghong Feng, **Noman Bashir**, Prashant Shenoy, David Irwin, and Beka Kosanovic. "Model-driven Per-panel Solar Anomaly Detection for Residential Arrays". In: ACM TCPS (2021).

[34] Pradeep Ambati, **Noman Bashir**, David Irwin, and Prashant Shenoy. "Waiting Game: Optimally Provisioning Fixed Resources for Cloud-Enabled Schedulers". In: ACM/IEEE SC. 2020.

[35] **Noman Bashir**, David Irwin, and Prashant Shenoy. "DeepSnow: Modeling the Impact of Snow on Solar Generation". In: ACM BuildSys. 2020.

[36] Menghong Feng, **Noman Bashir**, Prashant Shenoy, David Irwin, and Dragoljub Kosanovic. "SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays". In: ACM COMPASS. 2020.

[37] **Noman Bashir**, Dong Chen, David Irwin, and Prashant Shenoy. "Solar-TK: A Data-Driven Toolkit for Solar PV Performance Modeling and Forecasting". In: IEEE MASS. 2019.

[38] **Noman Bashir**, David Irwin, Prashant Shenoy, and Jay Taneja. "Mechanisms and Policies for Controlling Distributed Solar Capacity". In: ACM TOSN. 2018.

[39] **Noman Bashir**, David Irwin, Prashant Shenoy, and Jay Taneja. "Enforcing Fair Grid Energy Access for Controllable Distributed Solar Capacity". In: ACM BuildSys. 2017.

[40] **Noman Bashir**, Hira Shahzad Sardar, Mashood Nasir, Naveed Ul Hassan, and Hassan A. Khan. "Lifetime Maximization of Lead-Acid Batteries in Small Scale UPS and Distributed Generation Systems". In: IEEE PowerTech. 2017.

[41] **Noman Bashir**, Zohaib Sharani, Khushboo Qayyum, and Affan A. Syed. "Delivering Smart Load-shedding for Highly-stressed Grids". In: IEEE SmartGridComm. 2015.

[42] *Carbon Containers Software Prototype*. https://github.com/carbonfirst/CarbonContainers. 2023.

[43] *CarbonScaler Software Prototype*. https://github.com/umassos/CarbonScaler. 2024.

[44] *Product Attributes to Impact Algorithm*. https://paia.mit.edu/. 2024.

[45] *Ecovisor Software Prototype*. https://github.com/carbonfirst/Ecovisor. 2023.

[46] Prashant Shenoy, Andrew A Chien, David Irwin, Mohammad Hajiesmaili, Vivienne Sze, Mani Srivastava, Line Roald, Yuvraj Agarwal, Rick Adrion, Ramesh Sitaraman, Neena Thota, Priya Donti, Zico Kolter, Deepak Rajagopal, John Birge, Jimi Oke, and Ali Hortaçsu. *National Science Foundation Expeditions in Computing for Computational Decarbonization of Societal Infrastructures at Mesoscales*. https://codecexp.us/. (Accessed on October 8, 2024). May 2024.

[47] David Irwin, Prashant Shenoy, and Michael Zink. *CCRI: New: A Community Testbed for Designing Carbon-Efficient Cloud Applications*. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2213636. (Accessed on October 8, 2024). May 2022.

[48] Charles E. Leiserson, Neil C. Thompson, Joel S. Emer, Bradley C. Kuszmaul, Butler W. Lampson, Daniel Sanchez, and Tao B. Schardl. "There's plenty of room at the Top: What will drive computer performance after Moore's law?" In: *Science* (2020).

[49] Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. "Dark Silicon and the End of Multicore Scaling". In: *SIGARCH Comput. Archit. News* (2011).