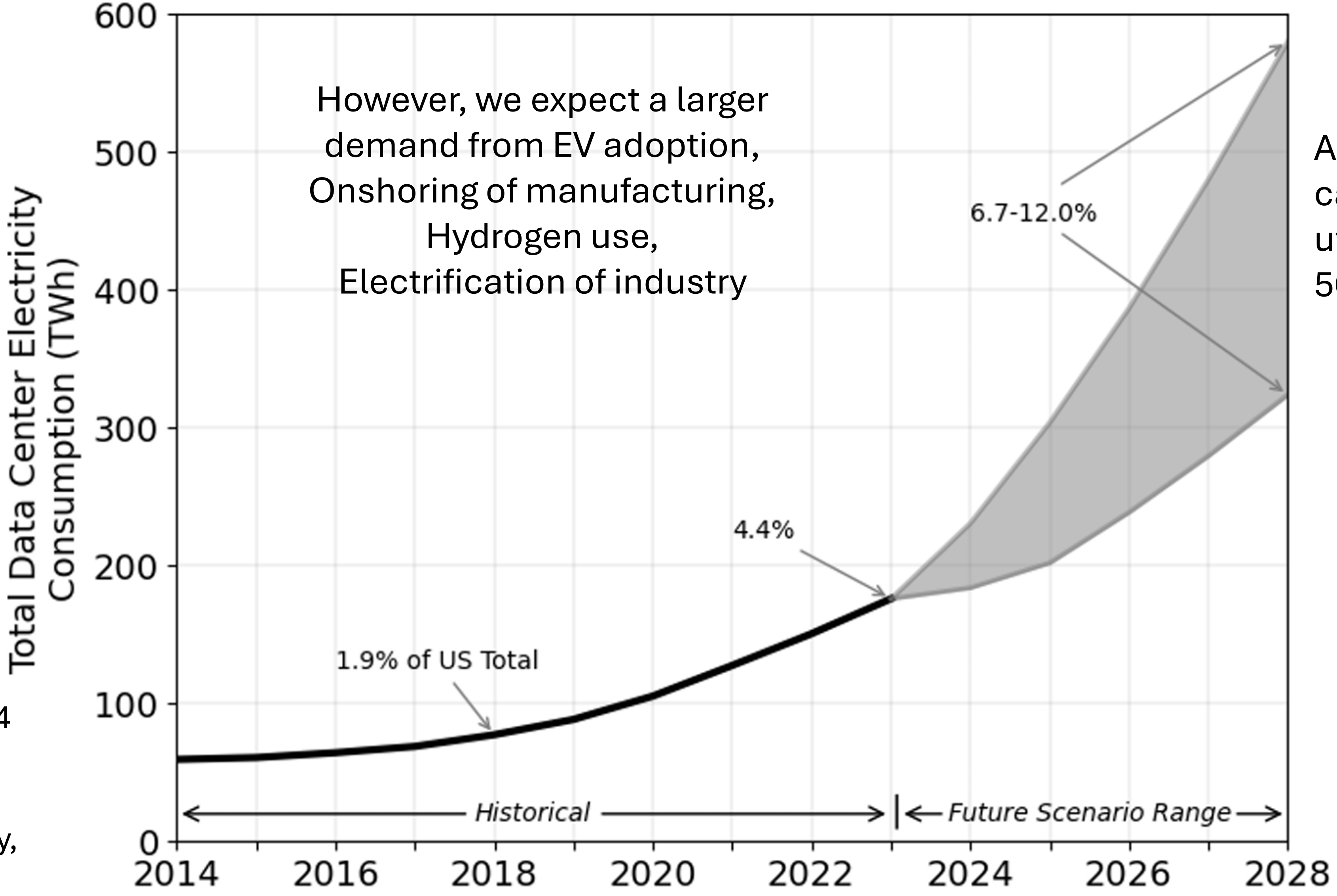
The Climate and Sustainability Implications of Generative AI

Tackling the Data Center Energy and Carbon Challenge,
Opportunity for whole-of-MIT collaboration in partnership with Industry

Link *chip design* to *workflow management* to *data center architecture* to *building footprint* to *power generation*

MIT Climate Project
MITei MIT Energy Initiative
MCSC
MIT Schwarzman College of Computing
rle AT MIT
MIT LINCOLN LABORATORY
MIT Climate Policy Center
MIT CEEPR Center for Energy and Environmental Policy Research
MIT Center for Real Estate
MIT School of Engineering
MGHPCC
MIT AI Hardware Program

Data centers consumed about 4.4 percent of total electricity in the United States in 2023, expected to grow…



Assuming capacity utilization of 50%

Needs:

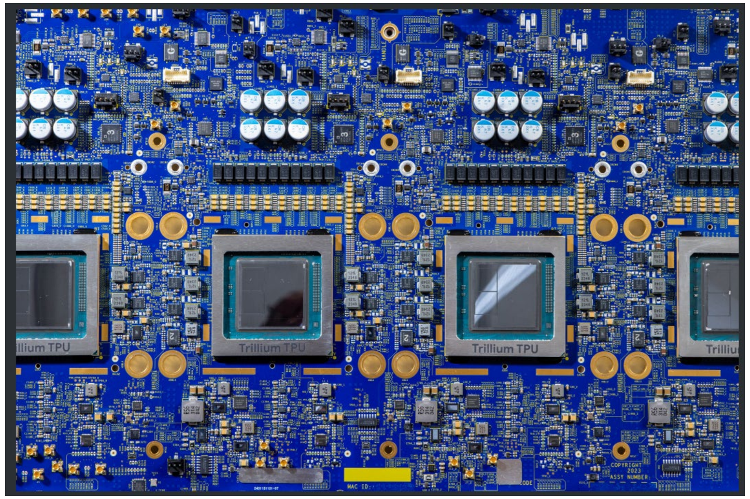
- Benchmarking
- Collaborations with utilities
- Technology development

Shehabi, A., et al. 2024 United States Data Center Energy Usage Report. LBNL, Berkeley, California.

Total AI energy demand highly uncertain – but the localized impact is indisputable

Google's first data center in 2006: The Dalles, OR cost $600m

https://datacenters.google/locations/the-dalles-oregon/



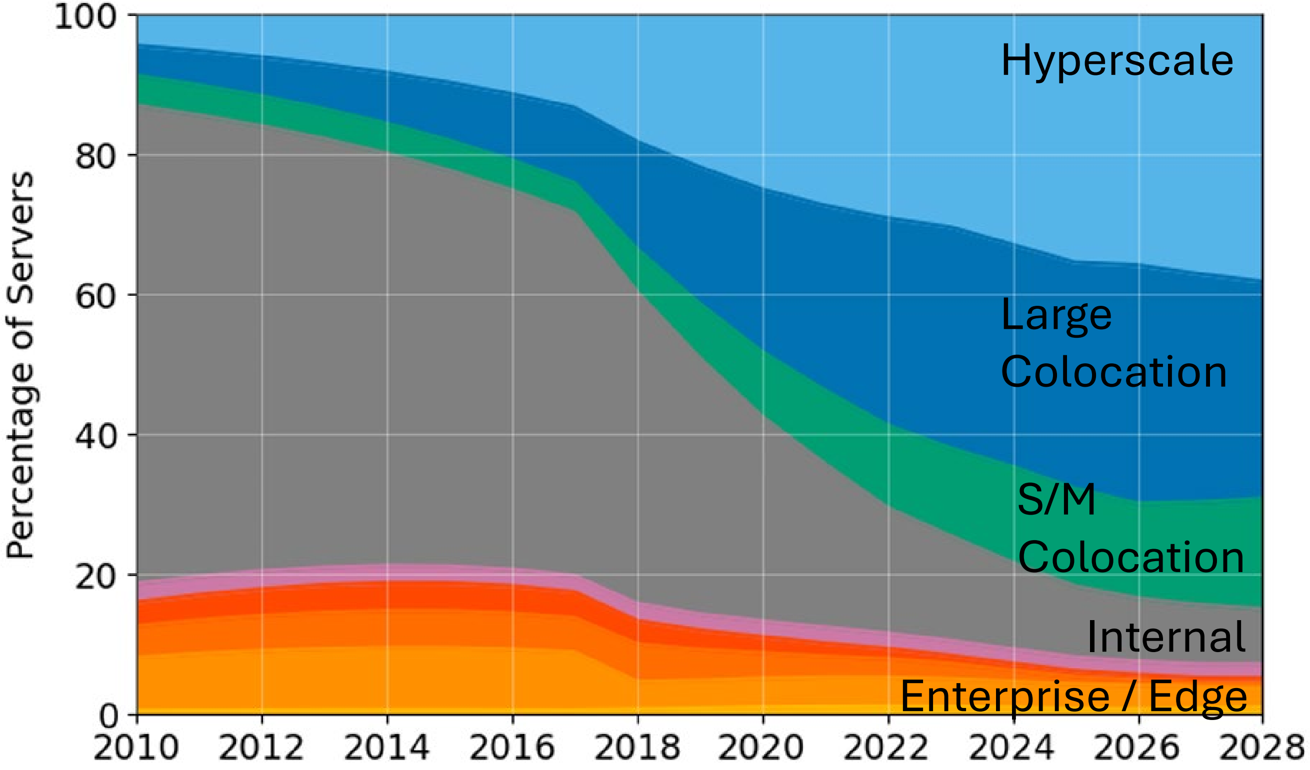
Typical CPU needs 250 to 500 watts to run, GPUs use up to 1,000 watts.

"Each the size of the Empire State Building, laid on its side across the desert"

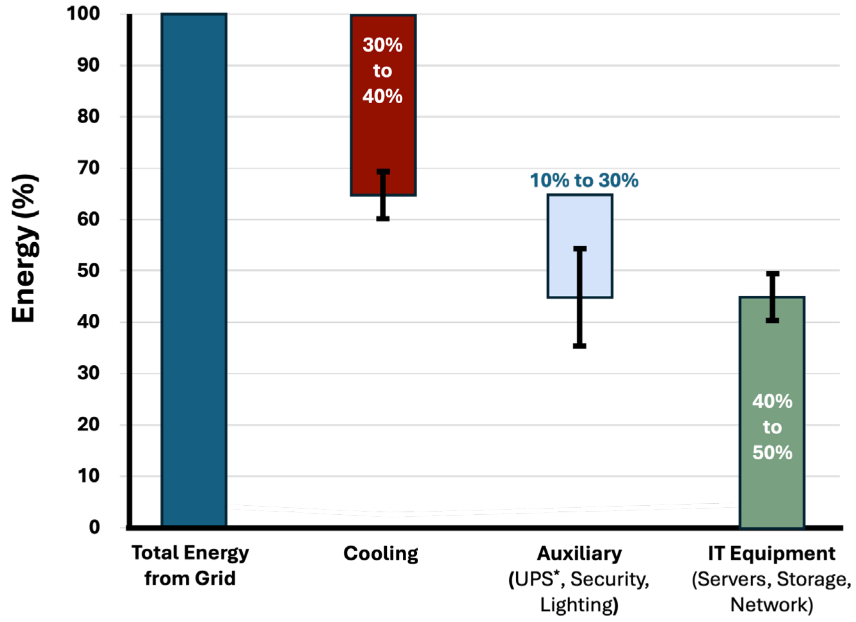
OpenAI announced plan to spend $100b in TX

https://www.costar.com/article/573467529/nations-first-stargate-data-center-in-west-texas-is-already-in-expansion-mode

Variation in server types lead to differences in processor types, cooling technologies and flexibility



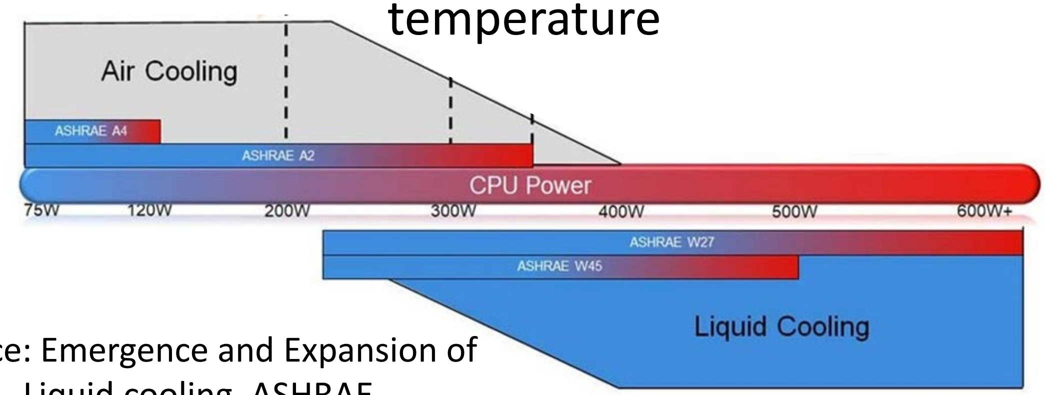
2024 United States Data Center Energy Usage Report, NREL



* UPS: Uninterruptible Power Supply

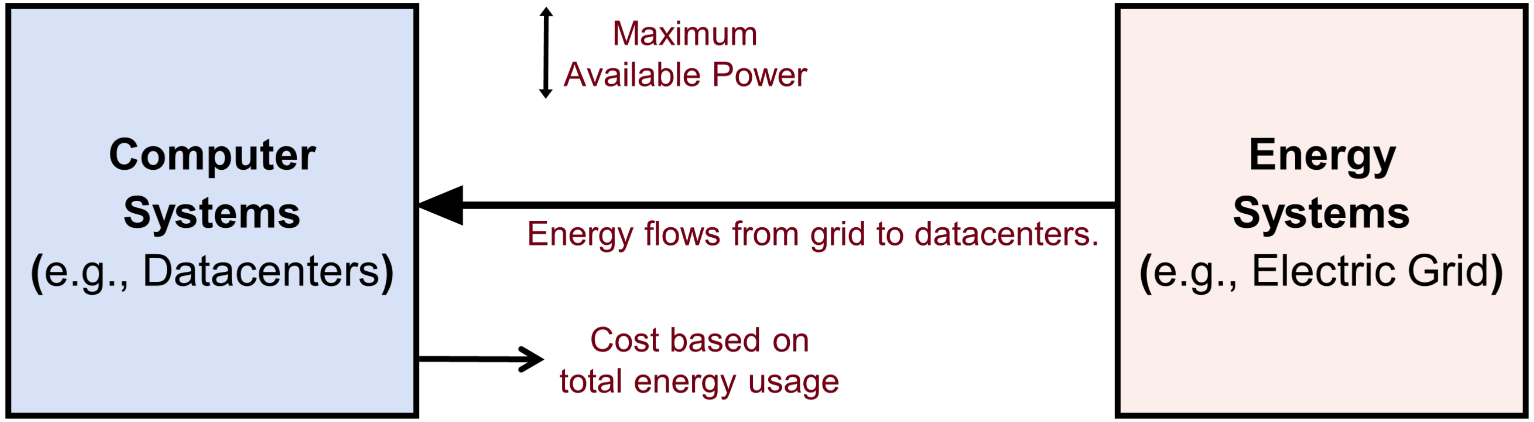
Aljbour, Jordan, Tom Wilson, and P. Patel. "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption." *EPRI White Paper no. 3002028905* (2024).

Air cooling versus liquid cooling, transition, and temperature



Source: Emergence and Expansion of Liquid cooling, ASHRAE

Current paradigm assumes unlimited and reliable power, we question whether this can or should persist



Increased demand & higher power intensity of AI workloads.

Large power swings in training clusters & at various timescales.

Using low-carbon energy introduces unique challenges.

Specialized hardware & variable lifetimes complicate operation.

Source: Noman Bashir

Lead sustainable, strategic technology (AI) deployment coupled with energy infrastructure modernization and decarbonization

Compute Demand: simulate & prototype chip to system for AI-driven workloads for hardware and software

Load & Operations Management: power electronics, time & space shifting

Market Design & Policy: Capacity expansion, cost, grid interconnection, rate & tariff design

Energy Supply: baseload via non-fossil technologies; control energy & storage cost and reliability

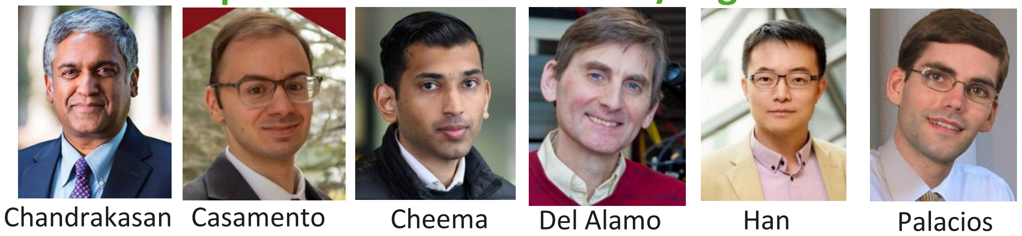
Metrics Design: Quantitative measures of how to inform possible actions at each layer of a computer system

Built Environment and Real Estate: air and liquid cooling, siting decisions, building systems and infrastructure

Need linked innovation in energy supply & compute demand to meet data center needs

Critical focus here:

Compute Demand: Hardware/Edge Devices



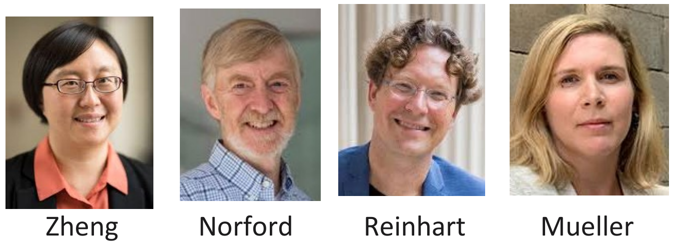
Compute Demand: Systems & Architecture

Energy Supply: Baseload Power

Grid Load and Operations

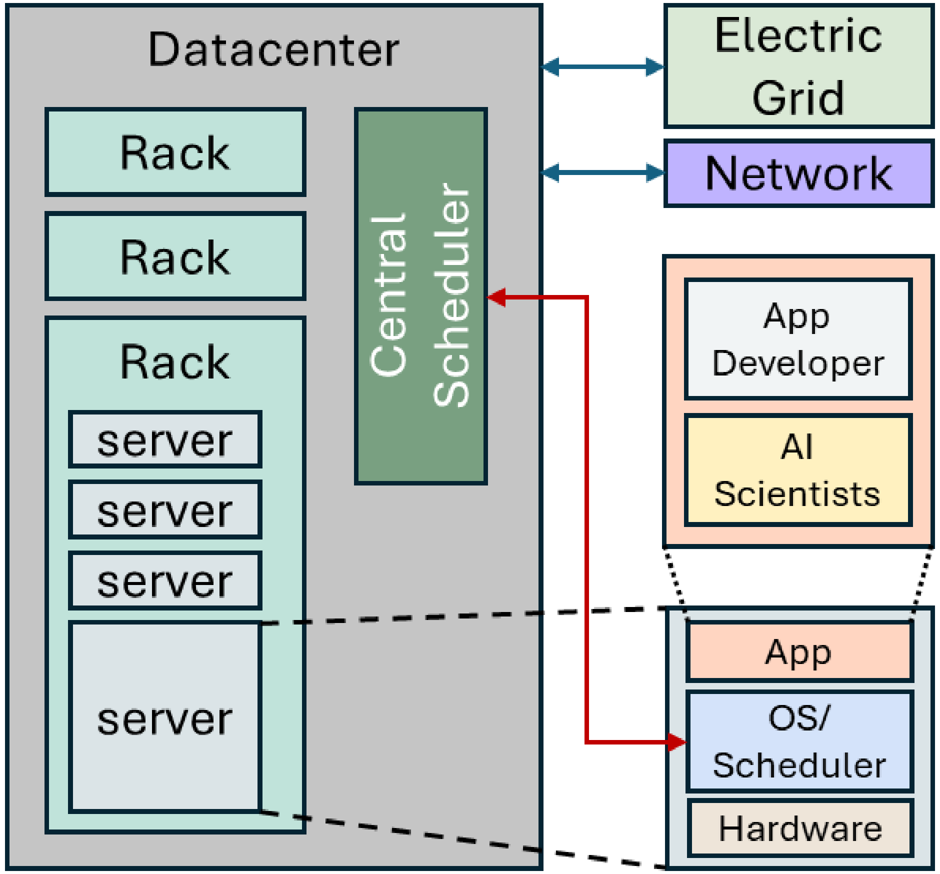
Market Design & Regulatory Policy

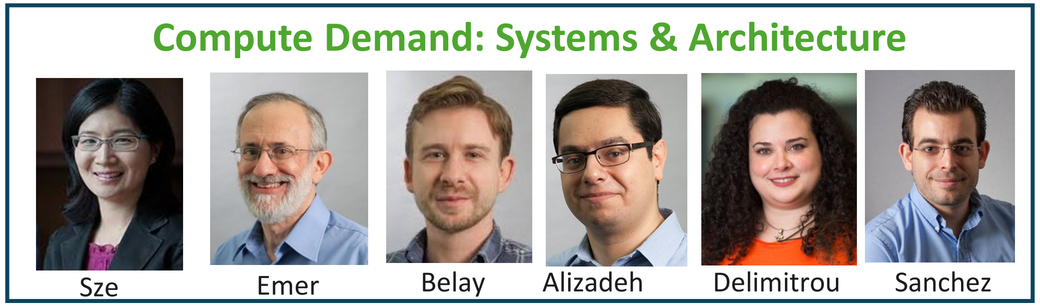
Real Estate & Building Energy Use

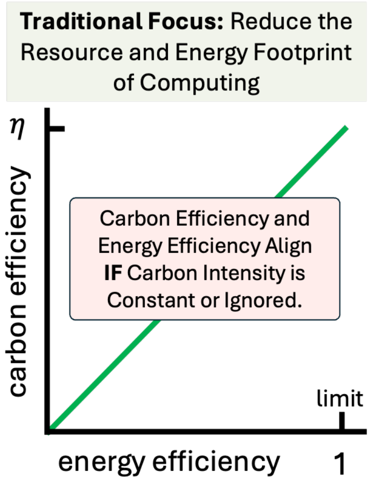


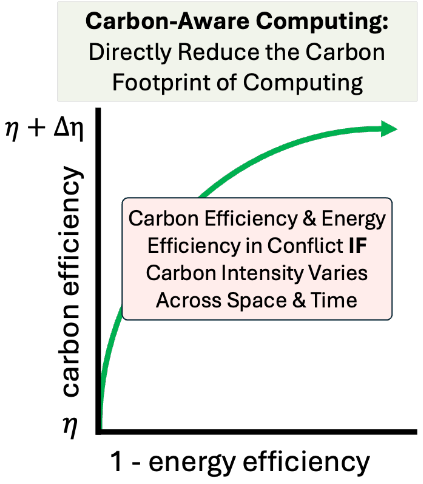
- Inform, verify & minimize risk for site selection and operational decisions
- Accelerate availability and realization of 24/7 carbon-free energy data center power solutions
- Inform policies, quantify economic benefits for grid interconnection, rate design, etc.

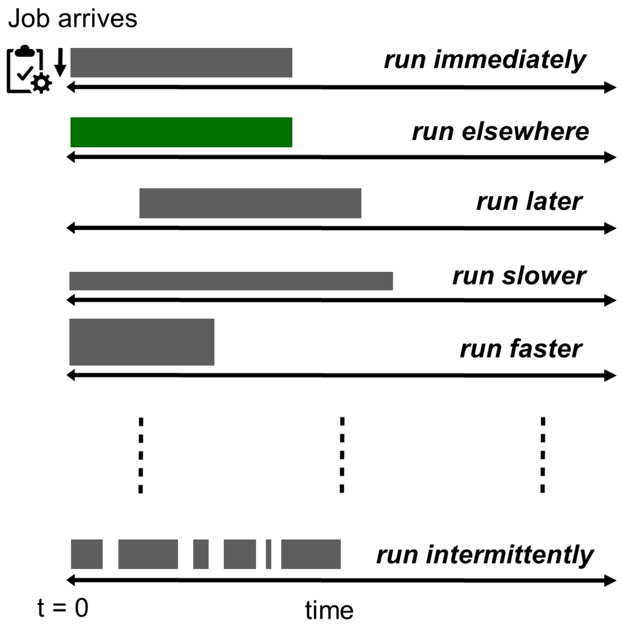
Compute Demand: Decisions at one level in the compute stack have ripple effects on energy, emissions, and performance throughout the system









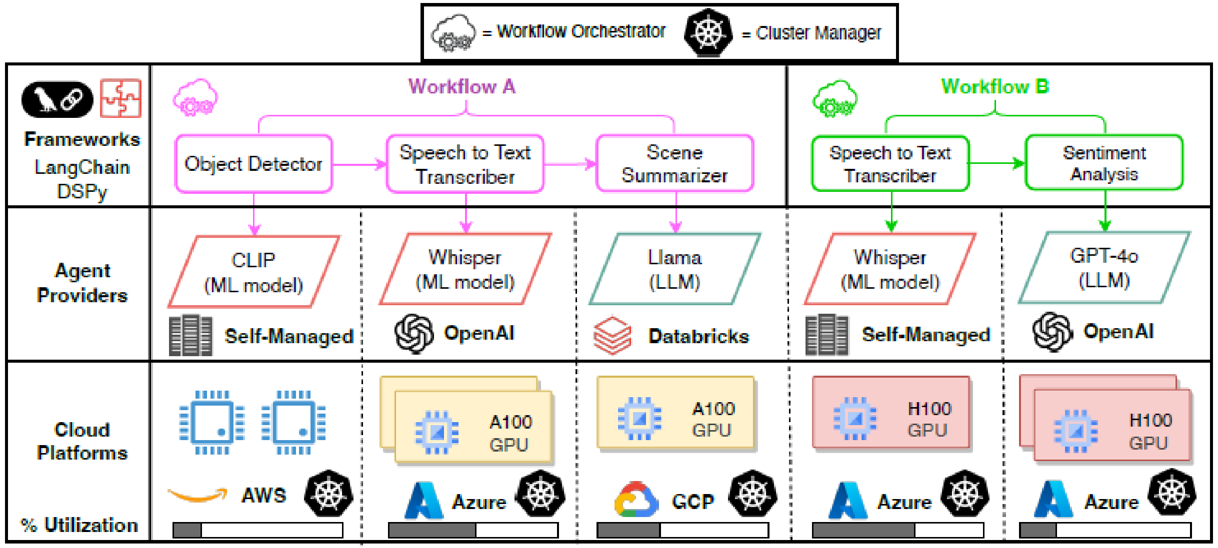


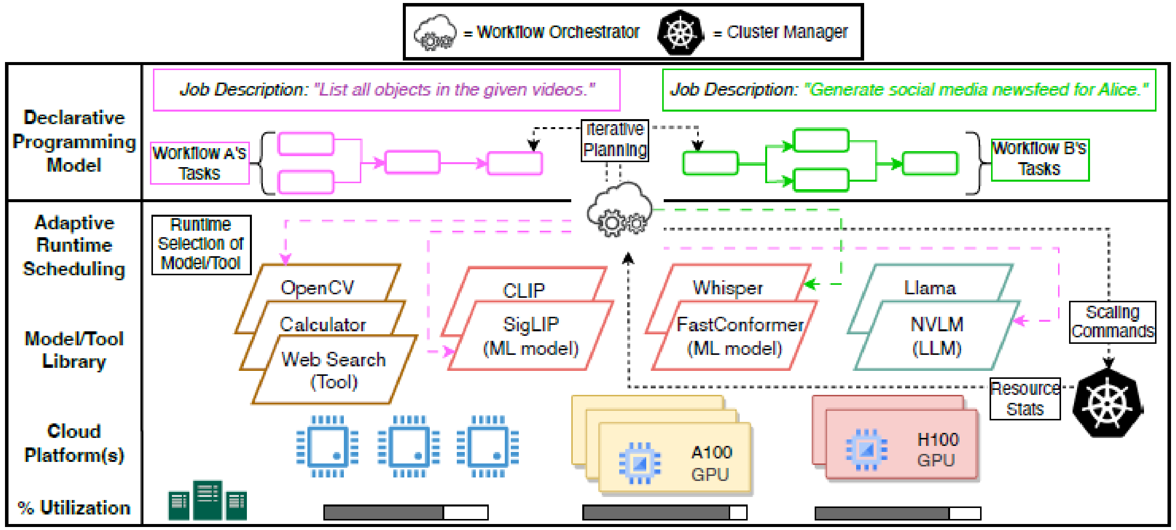
Source: Noman Bashir

Compute Demand: Today's AI systems tackle complex tasks using multiple interacting components, workflows grow ever deeper and self-improving

Adam Belay

Multiple frameworks, agents and providers!



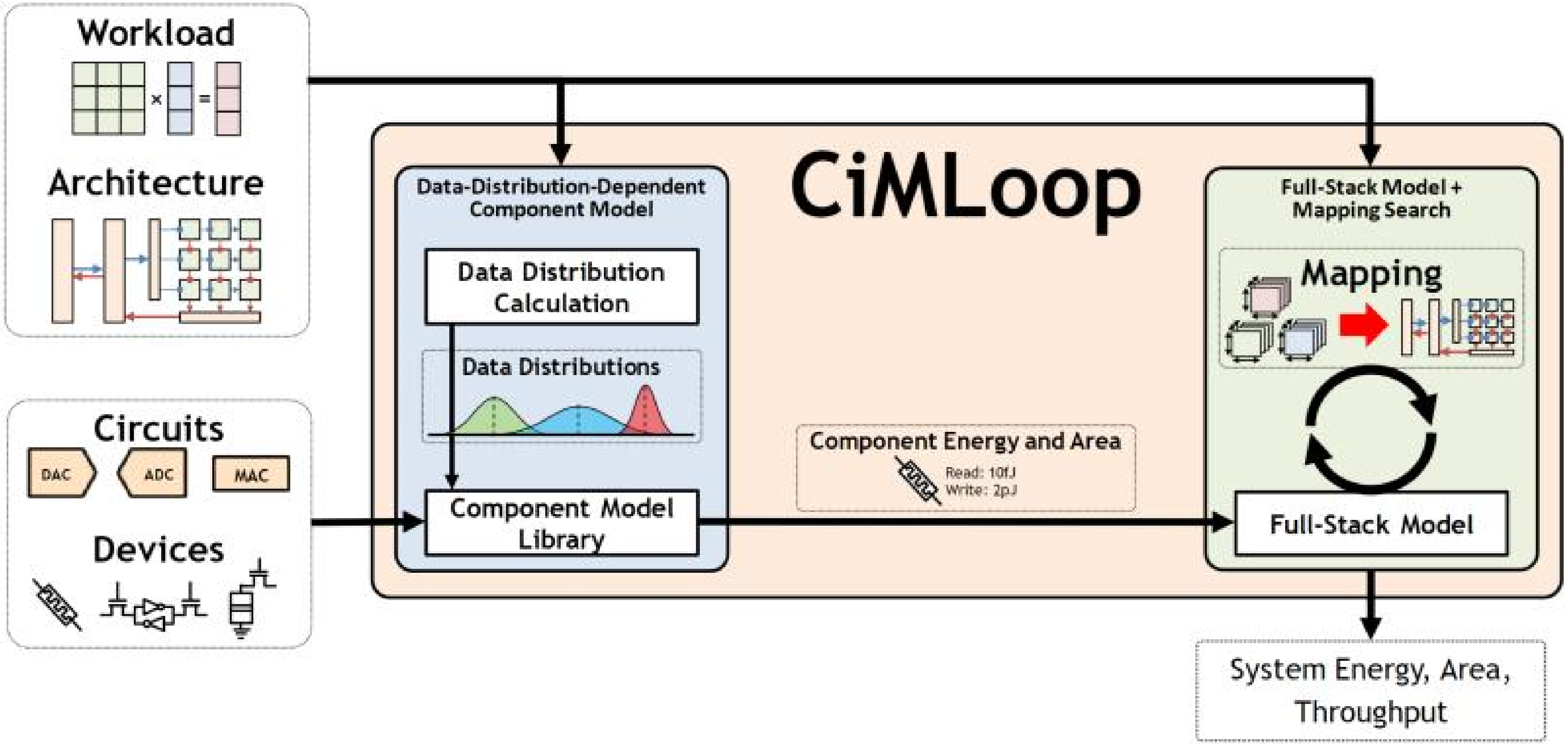


Demonstrates speedups up to ~ 3.4× in workflow completion times while delivering ~ 4.5× higher energy efficiency

https://arxiv.org/pdf/2501.16634

Compute Demand: Compute-in-Memory evaluate design choices at different levels of the stack, co-design across all levels, compare different implementations, and rapidly explore the design space.

Vivienne Sze Joel Emer

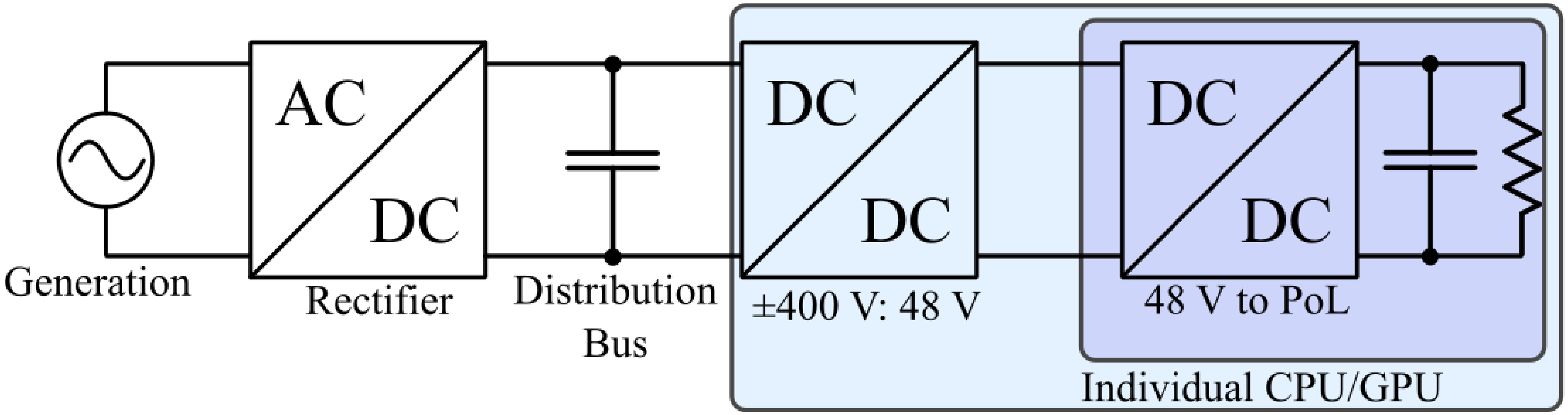


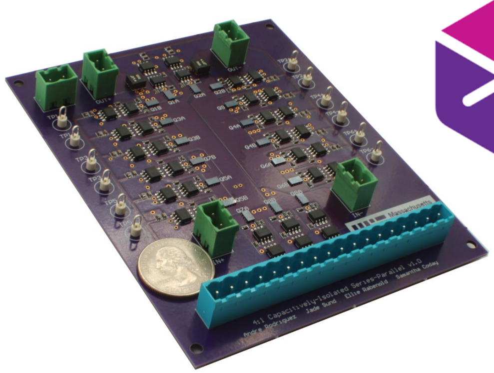
Bring all levels together in one model, bridge the device, circuits, and architecture research (& industry) communities

https://arxiv.org/pdf/2405.07259

Load Management: Data Center power delivery needs topologies capable of high (extreme) conversion ratios while maintaining high efficiency and power density

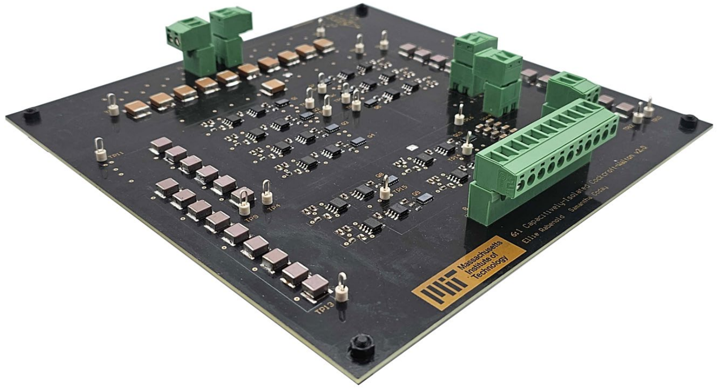
Samantha Coday





Capacitively-isolated series-parallel converter designed for ±400 V to 48 V.

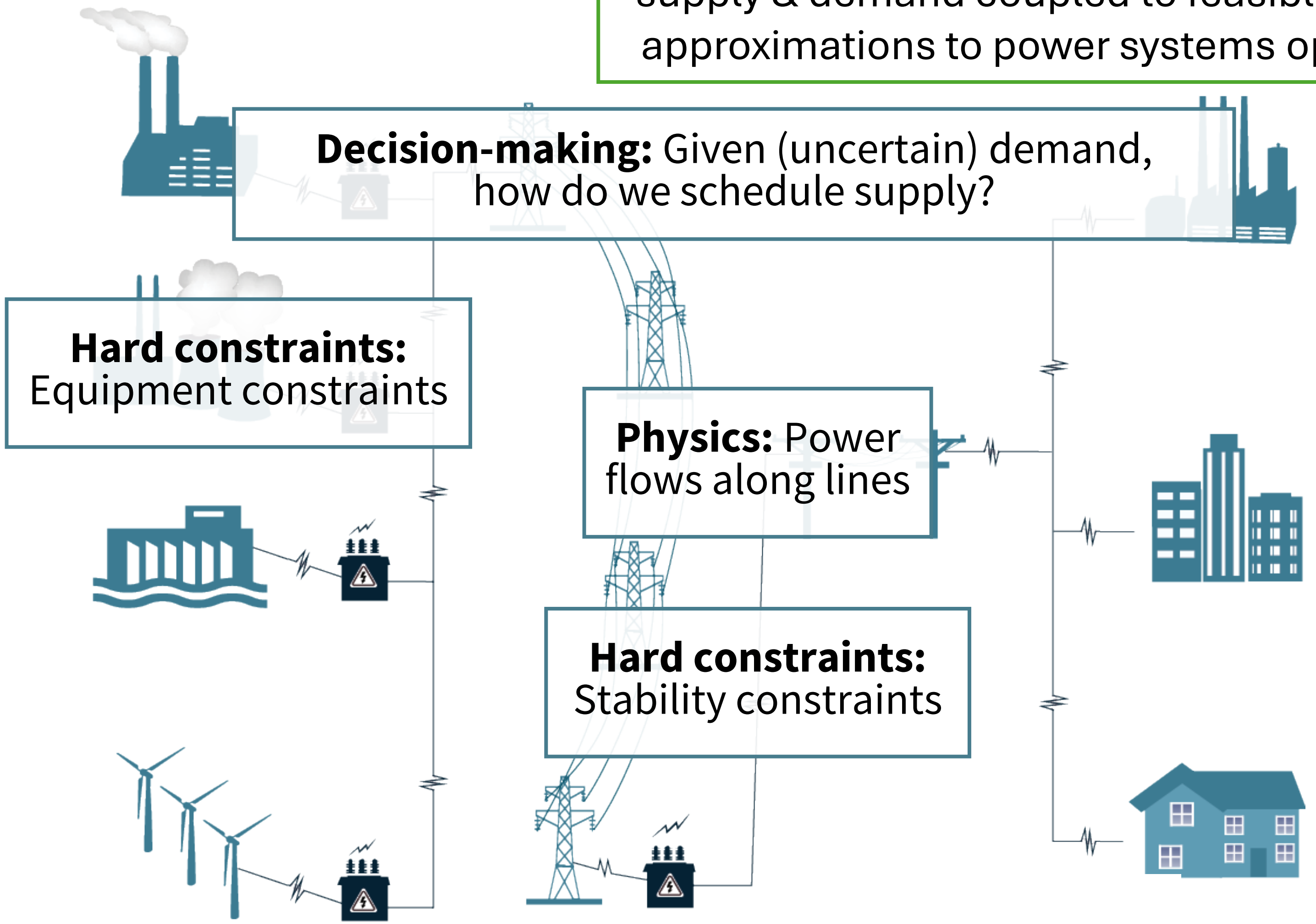
Designing converters which use (1) capacitive-based power conversion with energy-dense capacitors as the primary energy processing element (2) Partial power processing which allows for further modularity, decreased component stress and extreme conversion ratios (3) Input inductor converters which allow for better vertical integration and thermal performance.



A 6:1 capacitively-isolated Cockcroft-Walton converter used for PPP applications.

Load & Operations Management: Optimization-in-the-loop machine learning for power systems

Enabling decision-cognizant forecasting of supply & demand coupled to feasible, tractable approximations to power systems optimization



Priya Donti
MIT Schwarzman College of Computing

Trad. optimization & control
- Satisfies (many) constraints
- Struggles with speed / scale

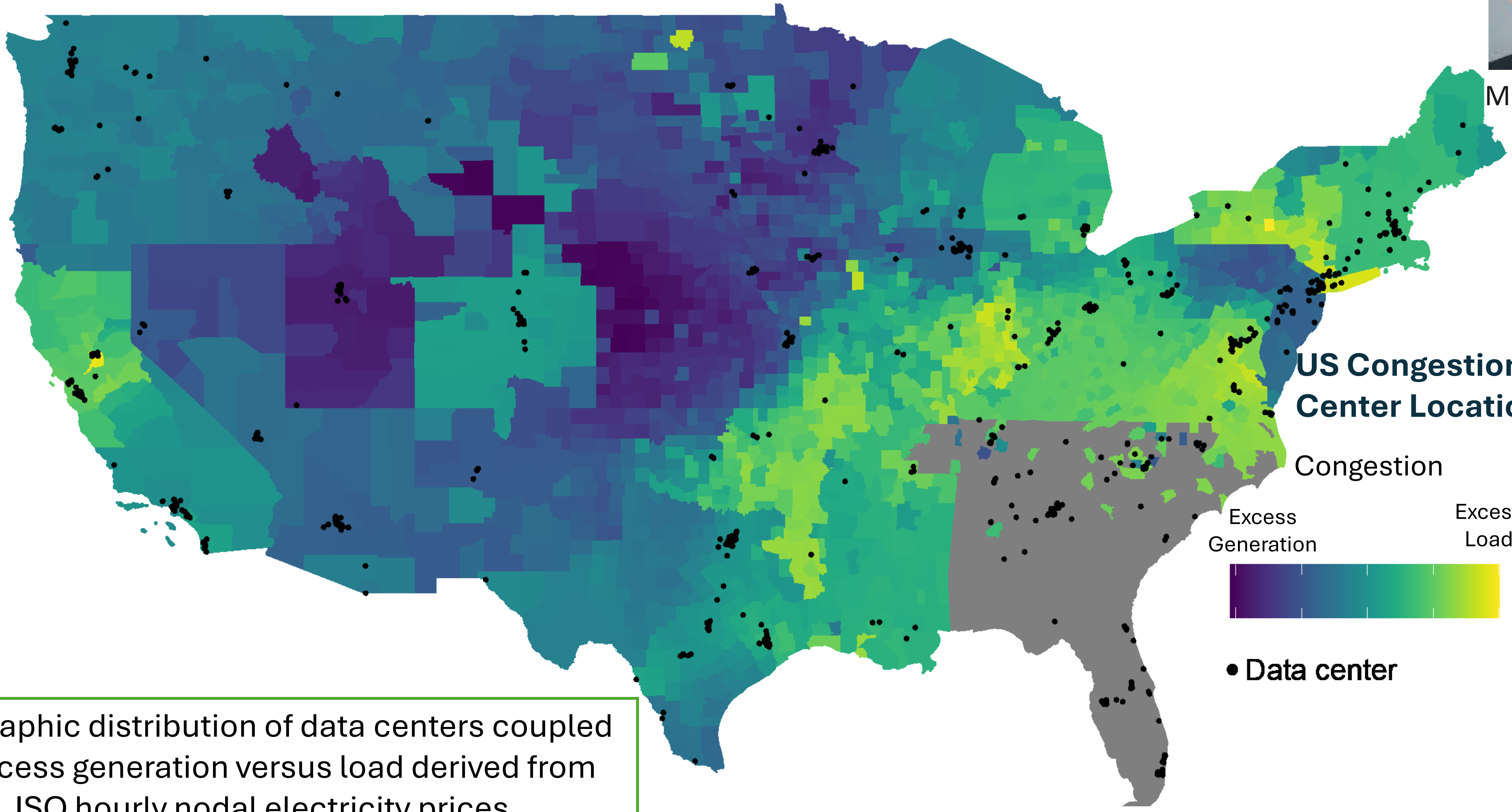
Machine learning (ML)
- Fast and scalable
- Struggles with constraints

Figure adapted from: US Congressional Budget Office
Source: Priya Donti

Market Design: Data Center Location driven by fiber network availability, electricity costs, reduce latency, customer proximity

Mert Demirer

MIT MANAGEMENT SLOAN SCHOOL



US Congestion & Data Center Locations (2022)

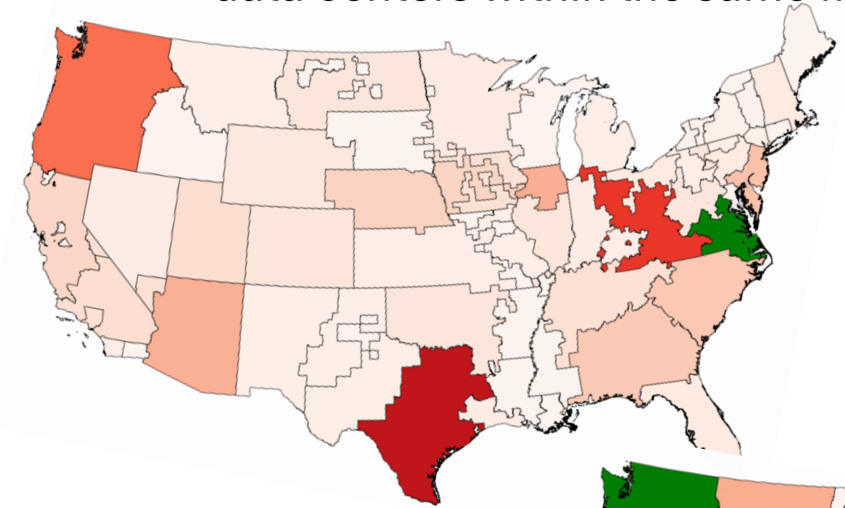
Geographic distribution of data centers coupled to excess generation versus load derived from ISO hourly nodal electricity prices

Market Design and Policy: Evaluating the Impact of Data Center Deployments on the Power System

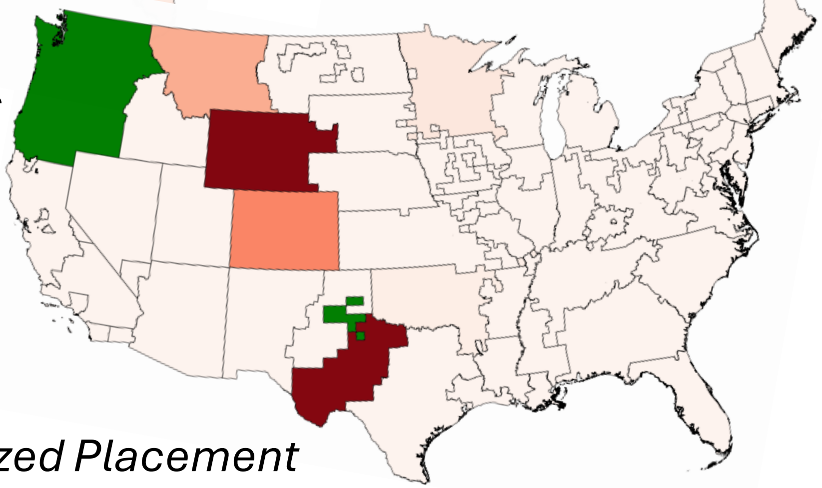
Chris Knittel

Geographic

Long run: placement of data centers
Short run: shifting computation burden across data centers within the same firm



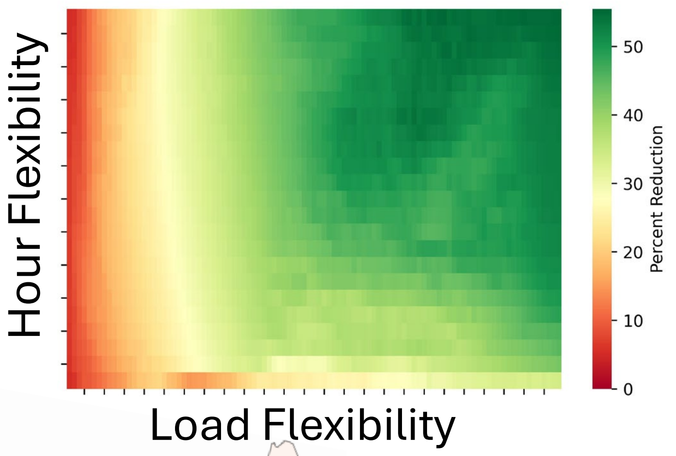
EPRI Data Center Forecasts

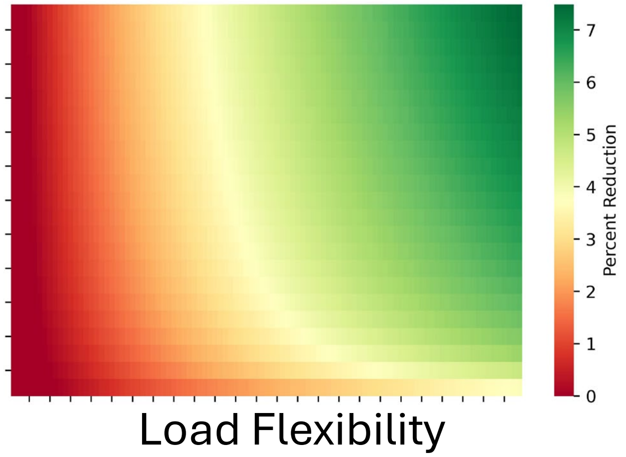


System Cost Optimized Placement

Temporal

Load reducing computation burden for a given task
Load shifting when a task is done



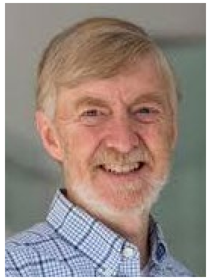


Grids need system-level policies informing locations of large pockets of load and optimal demand strategies for a cost-efficient transition to a net-zero power system

A single data center firm offering grid flexibility harmed in the marketplace; **Collective** flexibility improve consumer **&** firm outcomes

https://www.datacentermap.com/datacenters/

Building energy use: Multi-pronged approach including use of structural frame as an effective heat sink

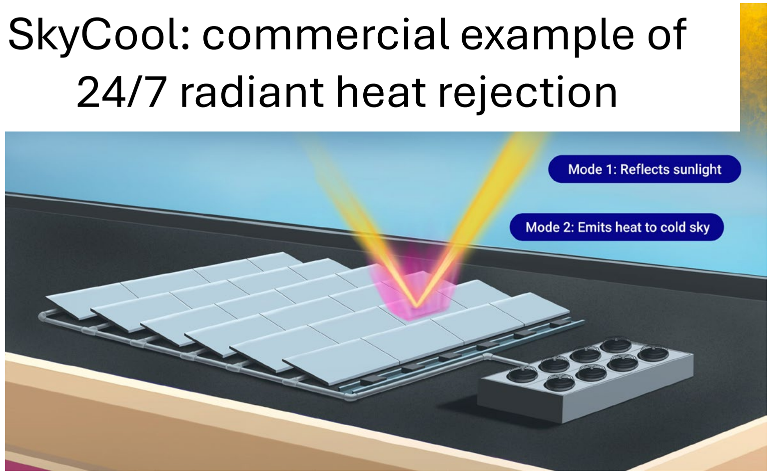


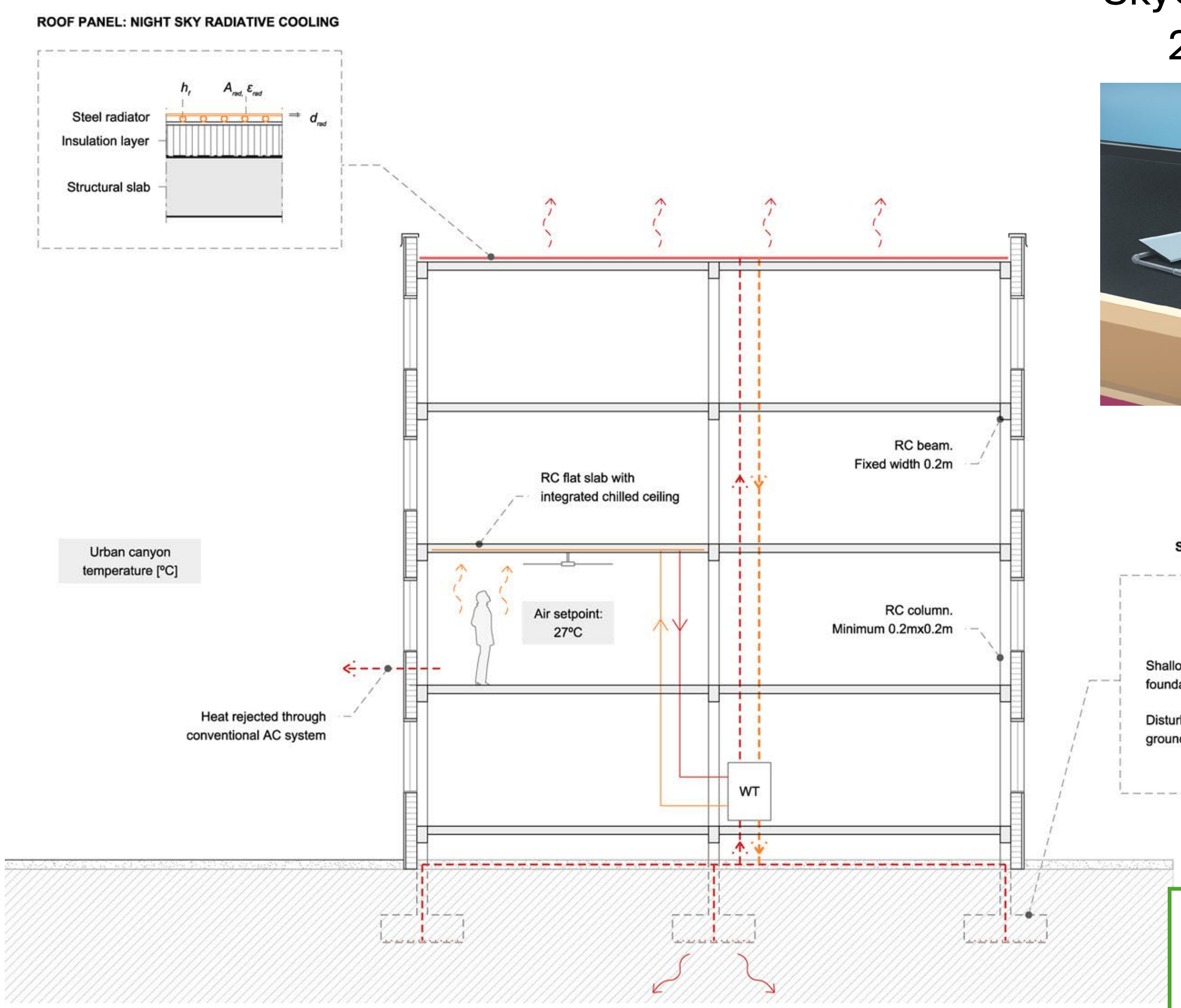
Les Norford

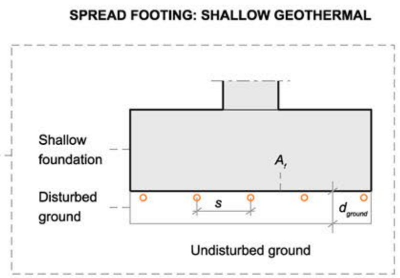
Caitlin Mueller

SA+P MIT SCHOOL OF ARCHITECTURE AND PLANNING

SkyCool: commercial example of 24/7 radiant heat rejection







Improve PUE by rejecting heat via low-wave radiation integrating heat dissipation systems within structural building components and coatings to reflect incident radiation

Leverage district heating and cooling systems and use heat for absorption cooling in warm weather

Energy Supply: Integrating design of data center and energy supply through localized grids powered by medium-to-large scale nuclear power stations

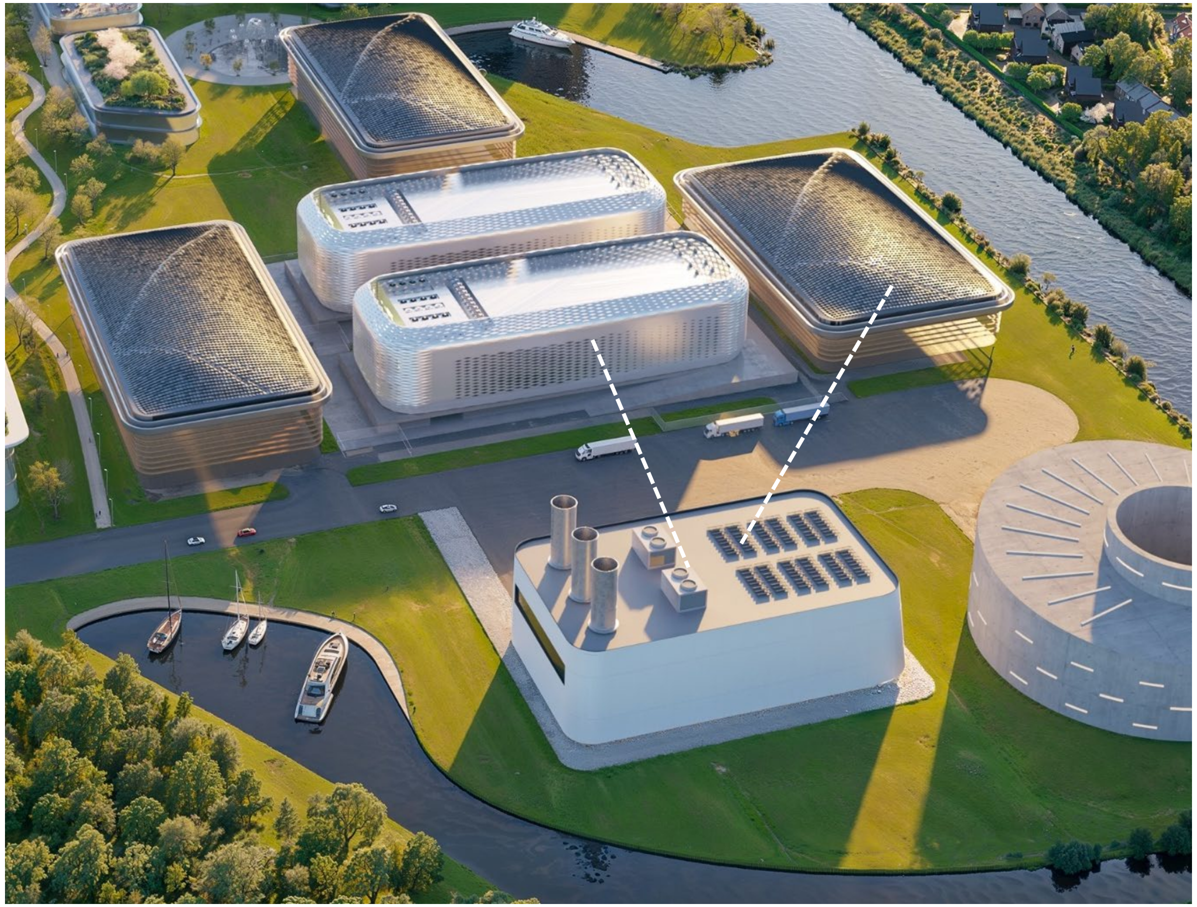
Emilio Baglietto

Iain Macdonald

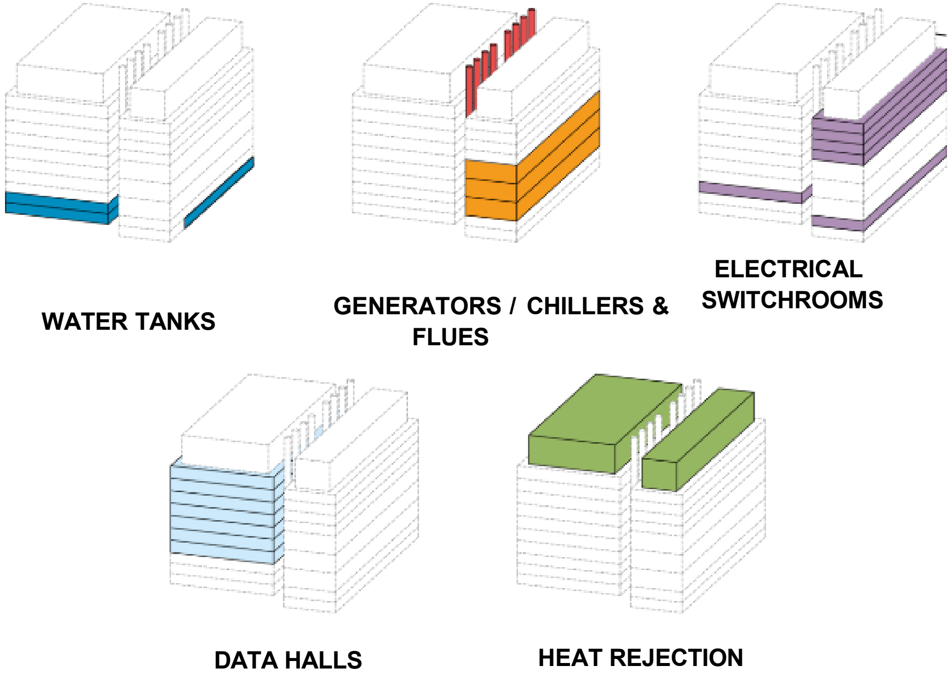
School of Engineering

Data Centers

Ai Infrastructure

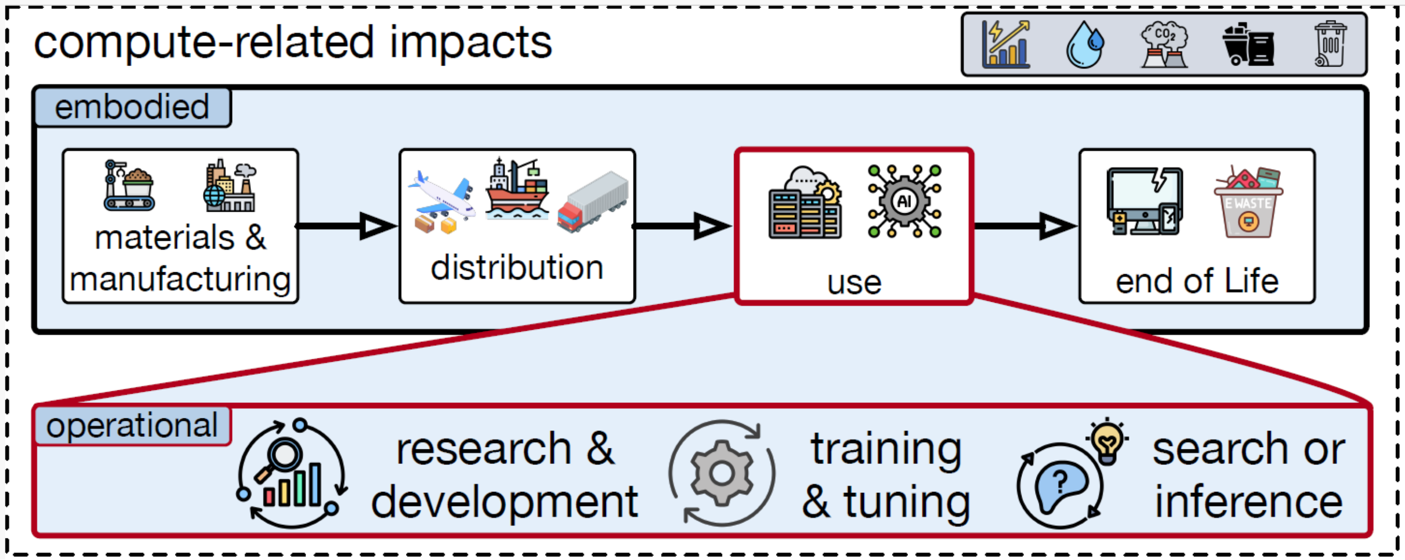


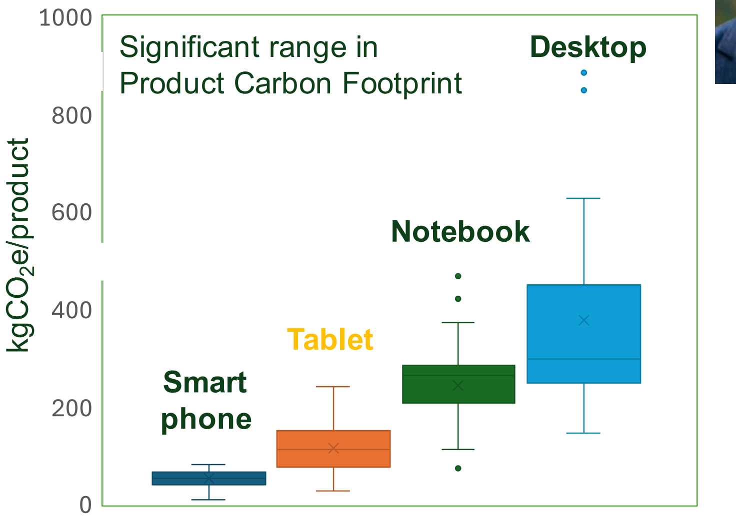
Advanced SMR+



High power density heat removal, innovations in liquid cooling, immersion cooling and forced air cooling, without the need for large water sources; Reduce power conversion infrastructure and minimize power transportation and distributions costs.

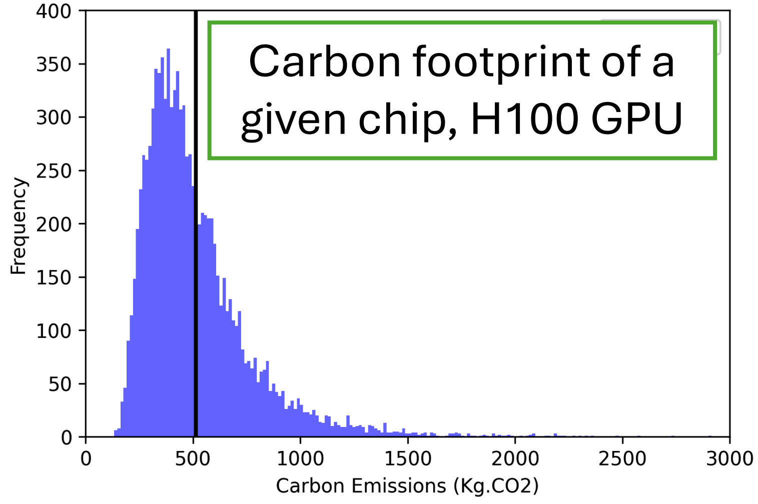
What's the impact of this system? Consider the full life cycle implications

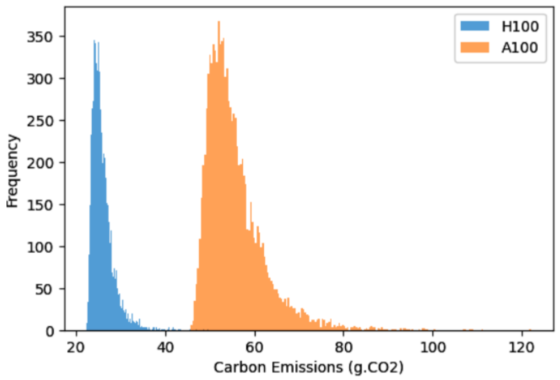




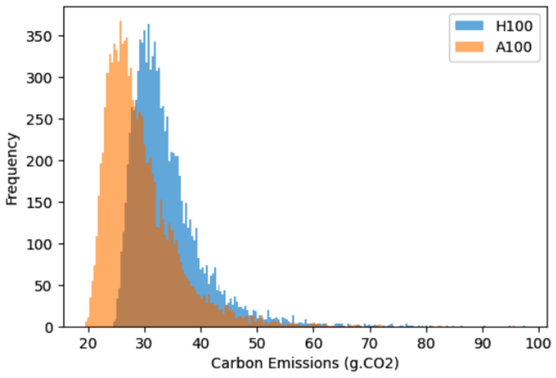
Lövehagen et al. *Renewable and Sustainable Energy Reviews*, 2023

Noman Bashir



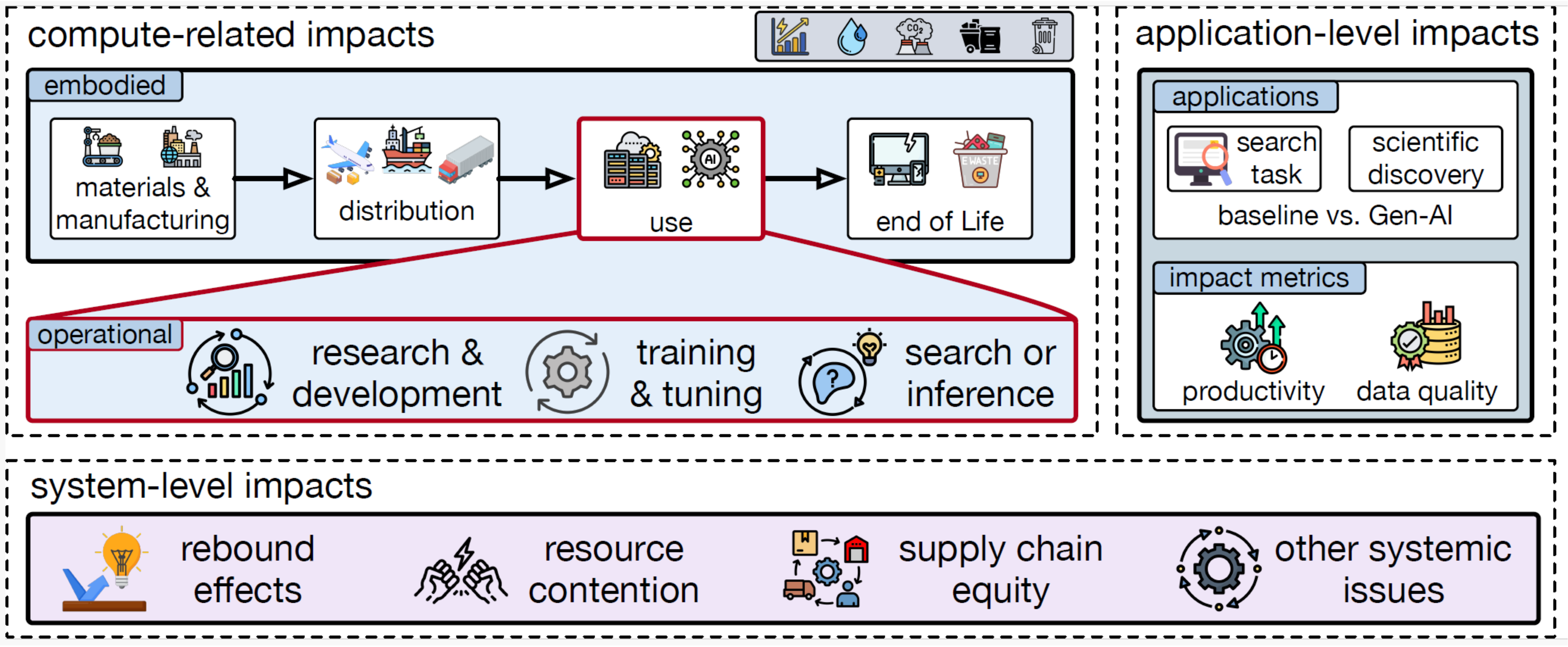


BERT
H100 reduces training time by 4x. **Net gain is 1.4x.**

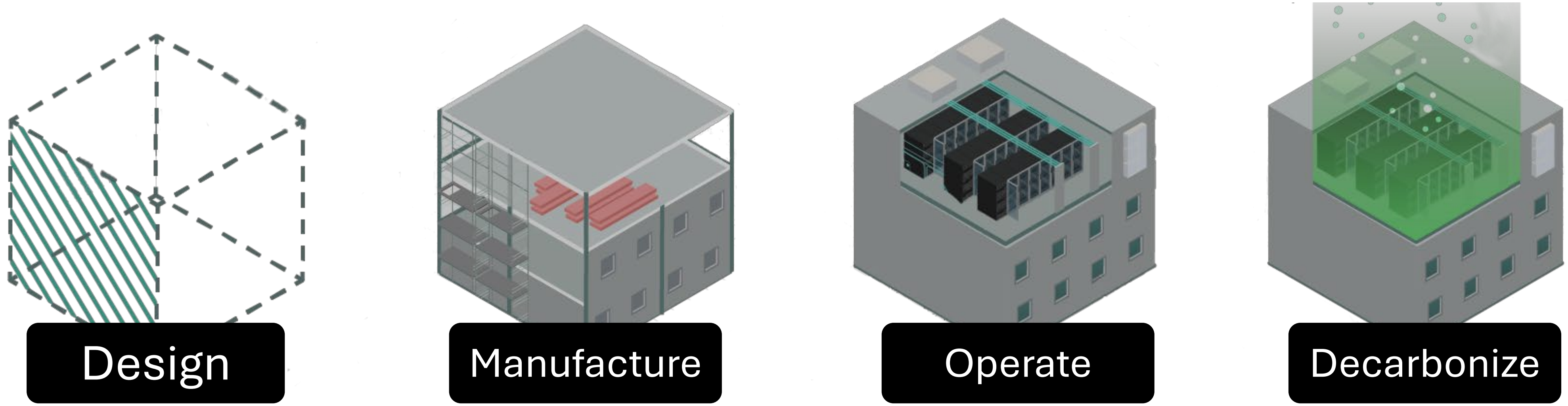


ResNet
H100 reduces training time by 1.5x. **No loss is 6%.**

Benefit–cost evaluation frameworks that encourage Gen-AI to develop beyond efficiency improvements to support social and environmental sustainability goals alongside economic opportunity



What are your challenges? Convene vital industry stakeholders across the value chain to address future technologies, systems and architectures



Chip Design

- Embodied vs operational tradeoff
- Handle power fluctuations in hardware

Resilient Workload Scheduling

- Model data movement costs
- Energy, carbon, performance, and reliability

Datacenter Architecture Design

- Location-specific cooling, server types
- Provide grid reliability services

Datacenter Demand Response

- Respond to electric grid’s DR signal
- Provide grid reliability services

Source: Noman Bashir