My research builds energy-efficient, general-purpose, post-von Neumann computers. General-purpose computing is the most impactful technology of the last fifty years, and it is essential that its growth continues. Unfortunately, general-purpose computing is currently under threat. The future of general-purpose computing depends on *energy efficiency* — from IoT to datacenters, energy efficiency determines computers' capability, lifetime, and environmental impact.

The last decade saw a dramatic shift towards specialized hardware. Hardware specialization has delivered large gains in energy efficiency, but the loss of *general-purpose programmability in high-level software* threatens the long-term value of computing to society. *The challenge for the next decade is to reconcile programmability and energy efficiency.* The efficiency of general-purpose computers must be improved by orders of magnitude to enable truly *sustainable* growth of computing.

Computing is a major contributor to global unsustainability. The environmental waste from computing is already comparable to the global aviation industry, and it is growing rapidly. My research builds computer systems with *sustainability* as a first-order design objective. This means designing systems to minimize their operating power (to reduce energy consumption) and maximize their lifetime (to amortize environmental impact from manufacturing).

Improving efficiency through hardware specialization is unsustainable and a looming crisis. Current general-purpose "von Neumann" processors are shockingly inefficient, *wasting 99% of their energy*. To improve efficiency, industry and academia have recently turned to specialized hardware accelerators. Accelerators' efficiency comes at steep cost, however. Specialization significantly increases environmental impact from manufacturing because it lowers system utilization and lifetime — by design, specialized hardware is only applicable to a few computations.

Worse, specialization risks locking tomorrow's programmers into today's computations, preventing breakthrough applications or innovations from seeing the light of day. Historically, the most impactful applications of computers have been *unanticipated* by their designers. General-purpose programmability is essential to computing's long-term value because, without it, the loss to society, measured in applications not invented, will be dire. *My mission is to make compute abundant and readily accessible to application programmers by making high-level software as efficient as specialized hardware.*

Data movement is the root cause of energy inefficiency in von Neumann computers. Moving data consumes orders-of-magnitude more energy than compute (e.g., nJ for off-chip memory vs. pJ for arithmetic), and the gap is growing. Unfortunately, von Neumann designs force large amounts of unnecessary data movement, as they place compute far away from data and hide all data movement from software. My research builds data-centric programming models and computer systems that *treat data movement as a first-class citizen*, so that hardware and software can work *together* to minimize data movement. My research fundamentally changes the relationship between software and data, unlocking a host of techniques to reduce data movement that are unachievable in current architectures.

My research builds general-purpose systems that minimize operating energy and maximize system lifetime. "Energy-minimal dataflow fabrics" are a new general-purpose compiler and architecture that eliminates the wasted energy in von Neumann processors. Instead of streaming instructions through a shared processor pipeline, we compile a program to a circuit where data flows directly between dependent operations. Energy-minimal architectures achieve unprecedented efficiency (1-2 TOPS/W) — e.g., efficient enough to run a smart camera continuously for *five years* on a single AA battery (vs. weeks today) — on software written in a typical, high-level language. "Polymorphic cache hierarchies" introduce data-centric hardware-software interfaces that give software control over data movement. Polymorphic cache hierarchies unlock new optimizations, impossible on von Neumann designs, that reduce data movement and improve energy efficiency on the most challenging workloads — e.g., by 5× on graph analytics. Finally, my datacenter flash caches reduce writes by an order of magnitude, extending device lifetime from 3 to 10 years and reducing their carbon impact by more than half.

Translating research into real-world impact. I have invested heavily in making my research real. In academia, we have not merely designed architectures, but also pushed them to three complete silicon prototypes thus far. In 2022, I co-founded *Efficient Computer* (https://www.efficient.computer) to commercialize energy-minimal dataflow fabrics, where I am the Chief Scientist leading technical strategy. Finally, our general-purpose datacenter cache has replaced dozens of special-purpose software caches at Meta, including for the Facebook social graph.

My research shows that systems can be both general-purpose *and* highly energy-efficient. Energy-minimal dataflow architectures are fully programmable in high-level languages, and can compile and run C programs at energy efficiency within 2× of equivalent, fully specialized hardware. Similarly, polymorphic cache hierarchies are general-purpose and easy to program, replacing specialized memory hardware with general-purpose, data-centric software. These designs show that the supposed hard tradeoff between generality and efficiency is a false choice. My research forges a sustainable growth path for general-purpose computer architectures, unlocking the full benefits of computing for decades to come.