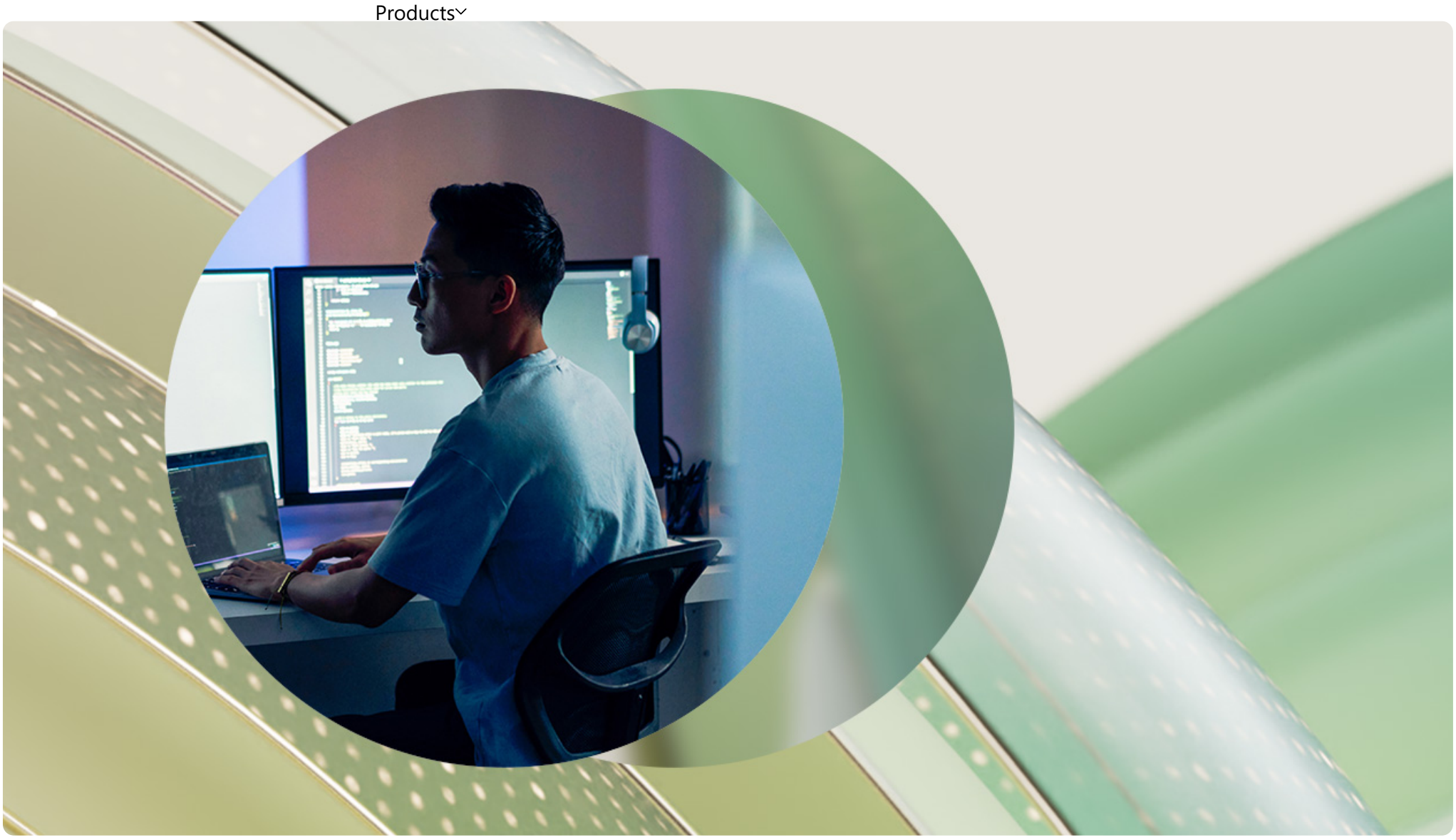


Thought leadership 4 minutes September 26, 2024

Sustainable by design: Innovating for energy efficiency in AI, part 2

By Mark Russinovich, CTO, Deputy CISO and Technical Fellow, Microsoft Azure

TAGS

AI Energy

Learn more about how we're making progress towards our sustainability commitments in part 1 of this blog: Sustainable by design: Innovating for energy efficiency in AI, part 1.

As we continue to deliver on our customer commitments to cloud and AI innovation, we remain resolute in our commitment to advancing sustainability. A critical part of achieving our company goal of becoming carbon negative by 2030 is reimagining our cloud and AI infrastructure with power and energy efficiency at the forefront.

We're pursuing our carbon negative goal through three primary pillars: carbon reduction, carbon-free electricity, and carbon removal. Within the pillar of carbon reduction, power efficiency and energy efficiency are fundamental to sustainability progress, for our company and for the industry as a whole.

Explore how we're advancing the sustainability of AI

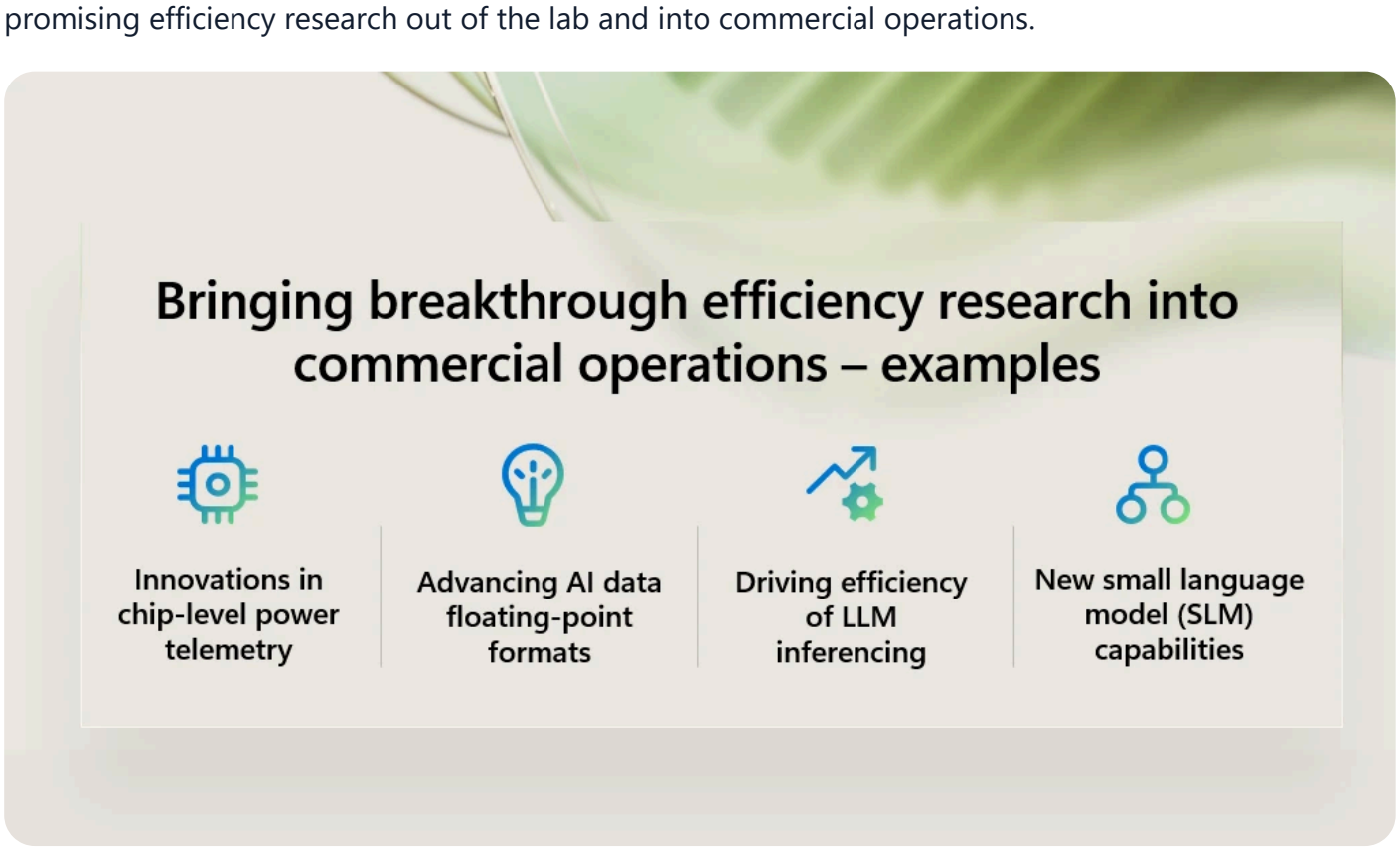
Explore our three areas of focus

Read more

Although the terms “power” and “energy” are generally used interchangeably, power efficiency has to do with managing peaks in power utilization, whereas energy efficiency has to do with reducing the overall amount of power consumed over time.

This distinction becomes important to the specifics of research and application because of the type of efficiency in play. For an example of energy efficiency, you might choose to explore small language models (SLMs) with fewer parameters that can run locally on your phone, using less overall processing power. To drive power efficiency, you might look for ways to improve the utilization of available power by improving predictions of workload requirements.

From datacenters to servers to silicon and throughout code, algorithms, and models, driving efficiency across a hyperscale cloud and AI infrastructure system comes down to optimizing the efficiency of every part of the system and how the system works as a whole. Many advances in efficiency have come from our research teams over the years, as we seek to explore bold new ideas and contribute to the global research community. In this blog, I’d like to share a few examples of how we’re bringing promising efficiency research out of the lab and into commercial operations.



Silicon-level power telemetry for accurate, real-time utilization data

We've made breakthroughs in delivering power telemetry down to the level of the silicon, providing a new level of precision in power management. Power telemetry on the chip uses firmware to help us understand the power profile of a workload while keeping the customer workload and data confidential. This informs the management software that provides an air traffic control service within the datacenter, allocating workloads to the most appropriate servers, processors, and storage resources to optimize efficiency.

Working collaboratively to advance industry standards for AI data formats

Inside the silicon, algorithms are working to solve problems by taking some input data, processing that data through a series of defined steps, and producing a result. Large language models (LLMs) are trained using machine learning algorithms that process vast amounts of data to learn patterns, relationships, and structures in language.

Simplified example from Microsoft Copilot*: Imagine teaching a child to write stories. The training algorithms are like the lessons and exercises you give the child. The model architecture is the child's brain, structured to understand and create stories. Inference algorithms are the child's thought process when writing a new story, and evaluation algorithms are the grades or feedback you give to improve their writing.*[1]

One of the ways to optimize algorithms for efficiency is to narrow the precision of floating-point data formats, which are specialized numerical representations used to handle real numbers efficiently. Working with the Open Compute Project, we've collaborated with other industry leaders to form the Microscaling Formats (MX) Alliance with the goal of creating and standardizing next-generation 6- and 4-bit data types for AI training and inferencing.

Narrower formats allow silicon to execute more efficient AI calculations per clock cycle, which accelerates model training and inference times. These models take up less space, which means they require fewer data fetches from memory, and can run with better performance and efficiency. Additionally, using fewer bits transfers less data over the interconnect, which can enhance application performance or cut network costs.

Driving efficiency of LLM inferencing through phase-splitting

Research also shows promise for novel approaches to large language model (LLM) inference, essentially separating the two phases of LLM inference onto separate machines, each well suited to that specific phase. Given the differences in the phases' resource needs, some machines can underclock their AI accelerators or even leverage older generation accelerators. Compared to current designs, this technique can deliver 2.35 times more throughput under the same power and cost budgets.[2]

Learn more and explore resources for AI efficiency

In addition to reimagining our own operations, we're working to empower developers and data scientists to build and optimize AI models that can achieve similar outcomes while requiring fewer resources. As mentioned earlier, small language models (SLMs) can provide a more efficient alternative to large language models (LLMs) for many use cases, such as fine-tuning experimentation on a variety of tasks or even grade school math problems.

In April 2024, we announced Phi-3, a family of open, highly capable, and cost-effective SLMs that outperform models of the same and larger sizes across a variety of language, reasoning, coding, and math benchmarks. This release expands the selection of high-quality models for customers, offering practical choices for composing and building generative AI applications. We then introduced new models to the Phi family, including Phi-3.5-MoE, a Mixture of Experts model that combines 16 smaller experts into one, and Phi-35-mini. Both of these models are multi-lingual, supporting more than 20 languages.

Learn more about how we're advancing sustainability through our Sustainable by design blog series, starting with Sustainable by design: Advancing the sustainability of AI.

[1]Excerpt from prompting Copilot with: please explain how algorithms relate to LLMs.

[2]Splitwise: Efficient generative LLM inference using phase splitting, Microsoft Research.

Related Posts

AI Dec 19 4 min read

Harnessing generative AI: The bold challenge and reward for industry leaders ›

Digital transformation Dec 16 4 min read

Seizing the AI opportunity: How to transform Canada's economy by 2030 >

Sustainability Dec 12 5 min read

3 ways AI is helping the planet >

Explore

Discover how the most trusted and comprehensive cloud can help you meet the challenges of a rapidly changing world.

Learn more about Microsoft Cloud solutions >

Connect with us on social