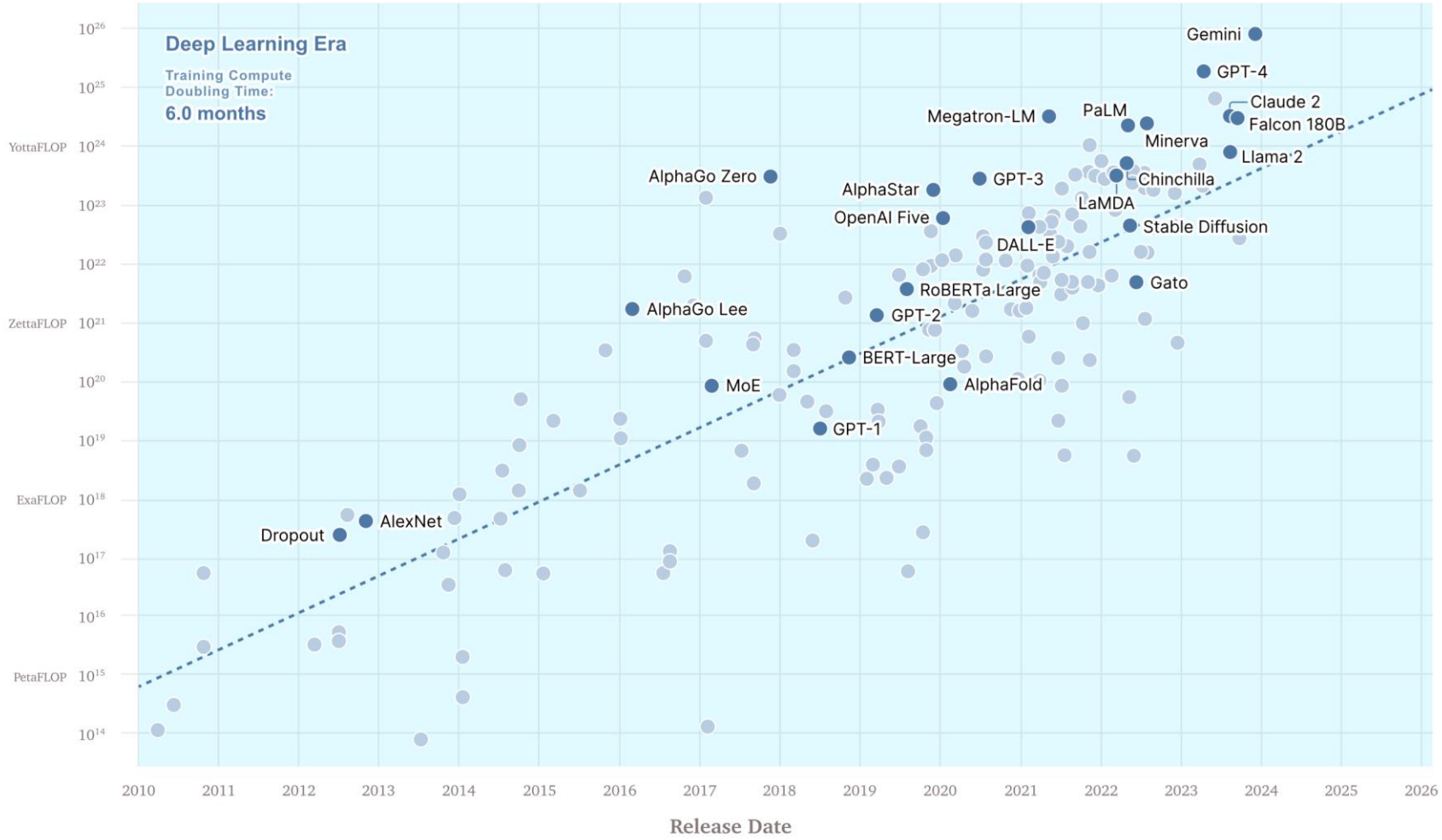
Towards Sustainable AI Infrastructure: Chips, Servers, and Racks

Jian Huang

Systems Platform Research Group

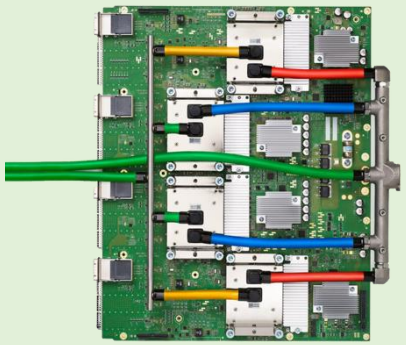
The Ever-Growing AI Demand for Computing Resources



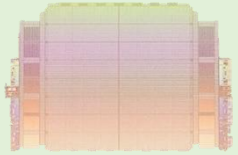
Computation (FLOPs) required for AI training doubles every 6 months

Source: Sastry, Girish et al. Computing Power and the Governance of Artificial Intelligence. 2024. arXiv:2402.08797.

AI Chips: The Hardware Backbone That Powers the AI Growth



Google TPU



Meta MTIA

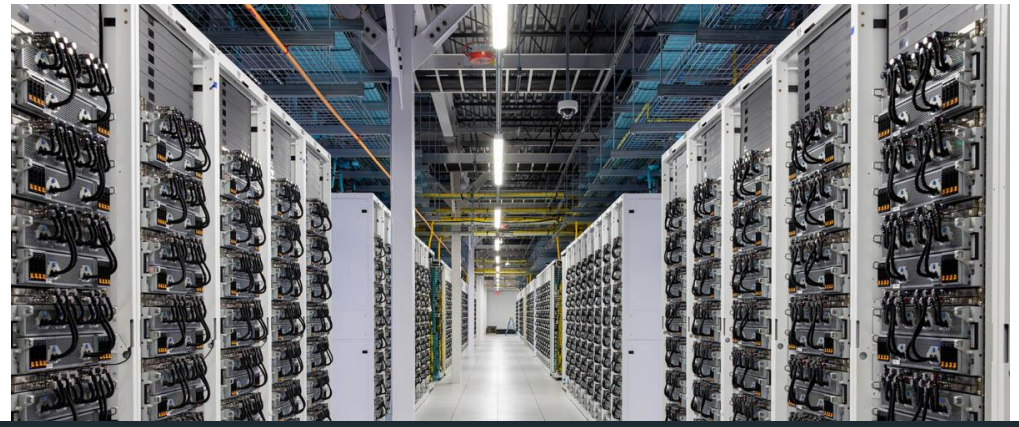
Graphcore IPU

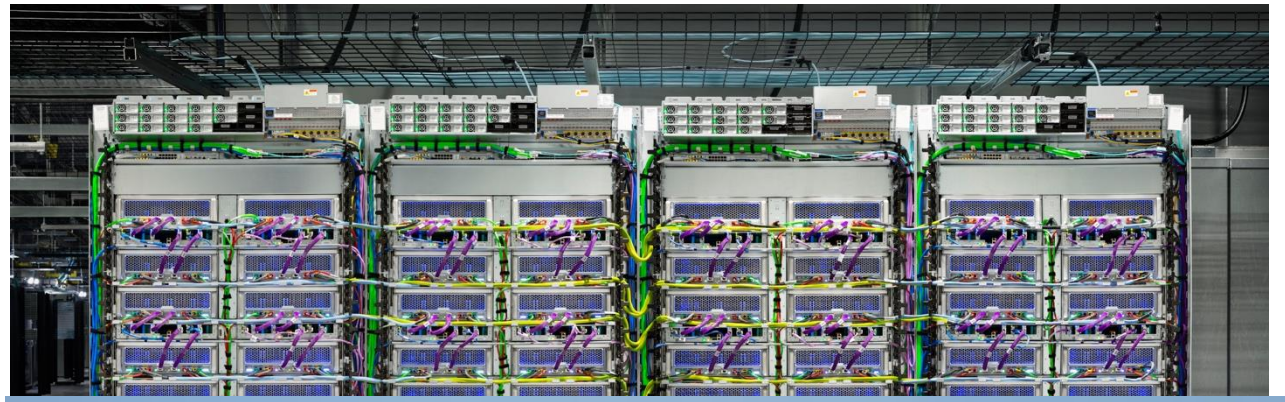
AWS Trainium

• • •

AI chips are widely deployed in data centers to power the ML workloads

AI Infrastructure: Scaling with Millions of AI Chips

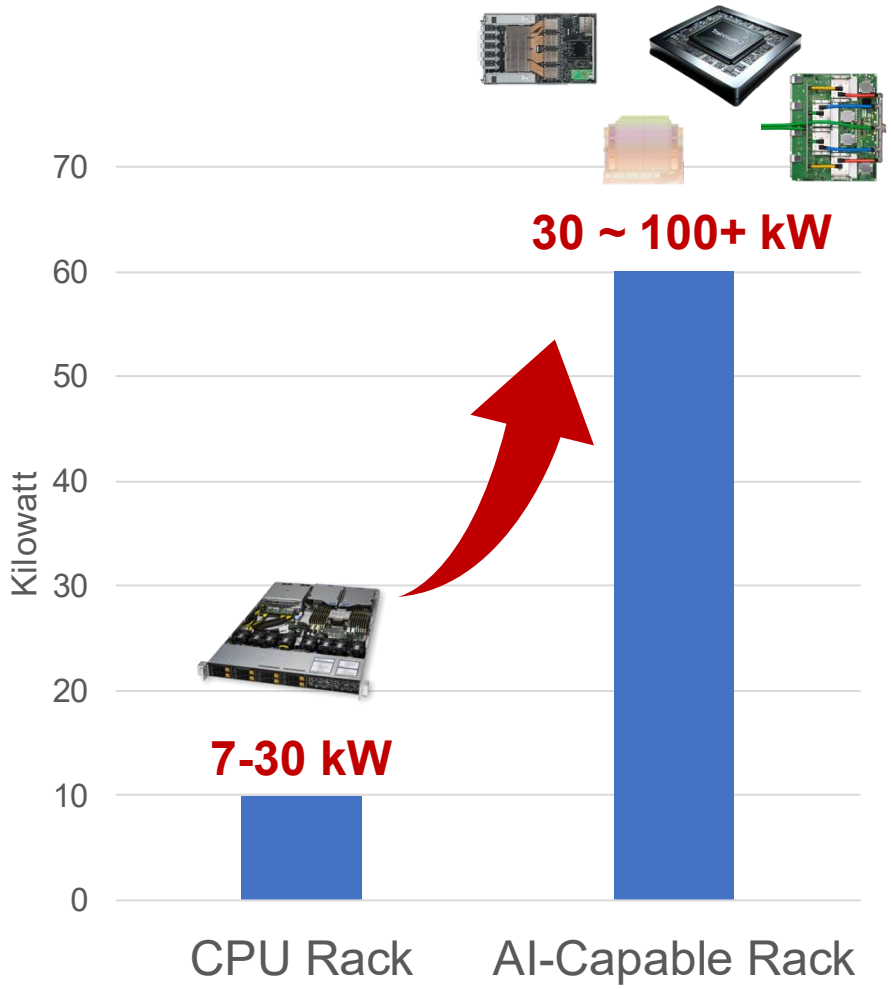




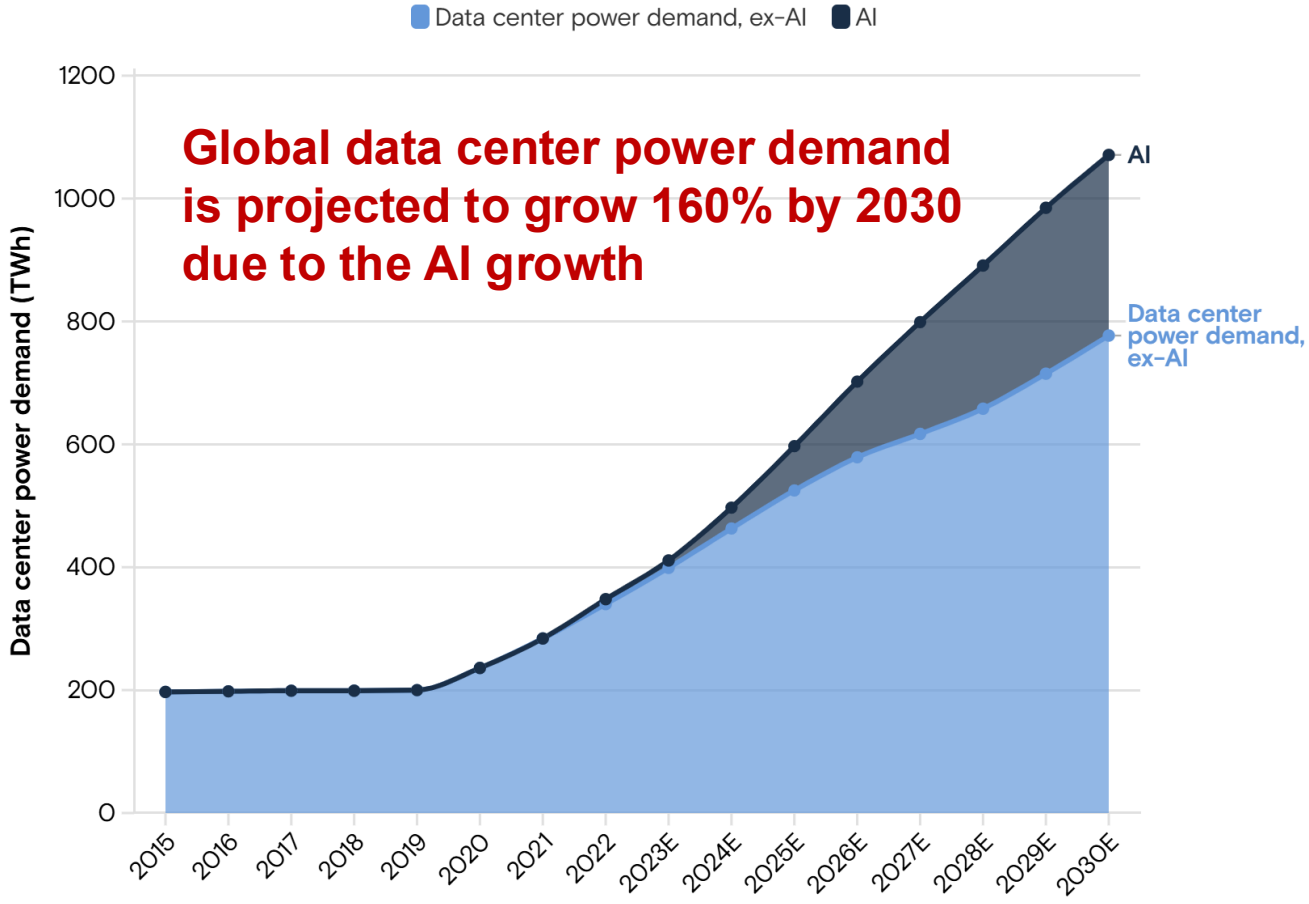
Meta's Mesa Data Center in Arizona

Amazon Data Center in Ohio

AI Infrastructures Are Power-Hungry

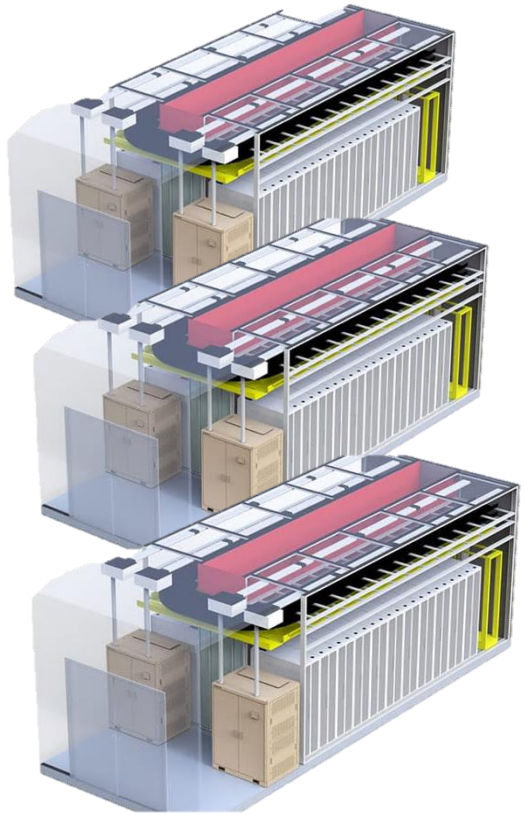


Source: Michael Wilson. 2025. Data Center Rack Power Costs: A Condensed Analysis



Source: Masanet et al. 2020. IEA, Cisco, Goldman Sachs Global Investment Research

Holistic Approach Towards Developing Sustainable AI Infrastructure



Energy/Compute Co-Design
(e.g., datacenter deployment with diverse renewable energy sources)

Energy/Carbon-aware System Software
(e.g., for scheduling tasks based on the carbon intensity of available power)

Server Component Reuse
(e.g., for hardware component reuse)

Energy-efficient AI Chips
(e.g., power gating and dynamic power management for AI chips)

Renewable Energy: A Promising Solution for Sustainable Data Centers

Microsoft will be carbon negative by 2030

Jan 16, 2020 | Brad Smith - President & Vice Chair

Google Sustainability Empowering individuals Working together Operating sustainably Reports

Overview Net-zero carbon Water stewardship Circular economy Nature & biodiversity Stories

IBM

IBM Newsroom News Media resources Inside IBM Blog

IBM Commits To Net Zero Greenhouse Gas Emissions By 2030

Feb 16, 2021

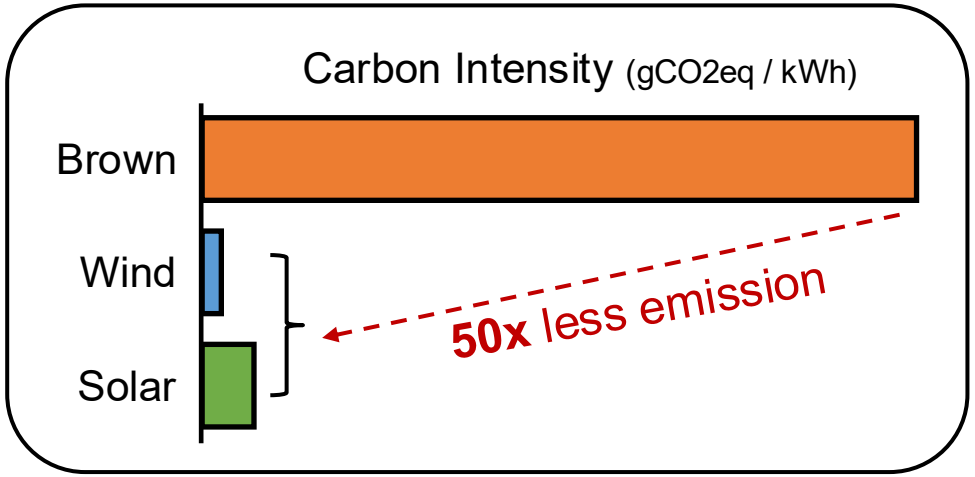
ARMONK, N.Y., Feb. 16, 2021 /PRNewswire/ -- IBM (NYSE: IBM) today announced that it will achieve net zero greenhouse gas emissions by 2030 to further its decades-long work to address the global climate crisis. The company will accomplish this goal by prioritizing actual reductions in its emissions, energy efficiency efforts and increased clean energy use across the more than 175 countries where it operates.

"I am proud that IBM is leading the way by taking actions to significantly reduce emissions," said Arvind Krishna, Chairman and Chief Executive Officer, IBM. "The climate crisis is one of the most pressing issues of our time. IBM's net zero pledge is a bold step forward that strengthens our long-standing climate leadership and positions our company years ahead of the targets set out in the Paris Climate Agreement."

To achieve its net zero goal IBM will:

Microsoft President Brad Smith, Chief Financial Officer Amy Hood and CEO Satya Nadella preparing to announce Microsoft's plan to be carbon negative by 2030. (Jan. 15, 2020/Photo by Brian Smale)

Major cloud providers aim to achieve net zero or carbon negative by 2030

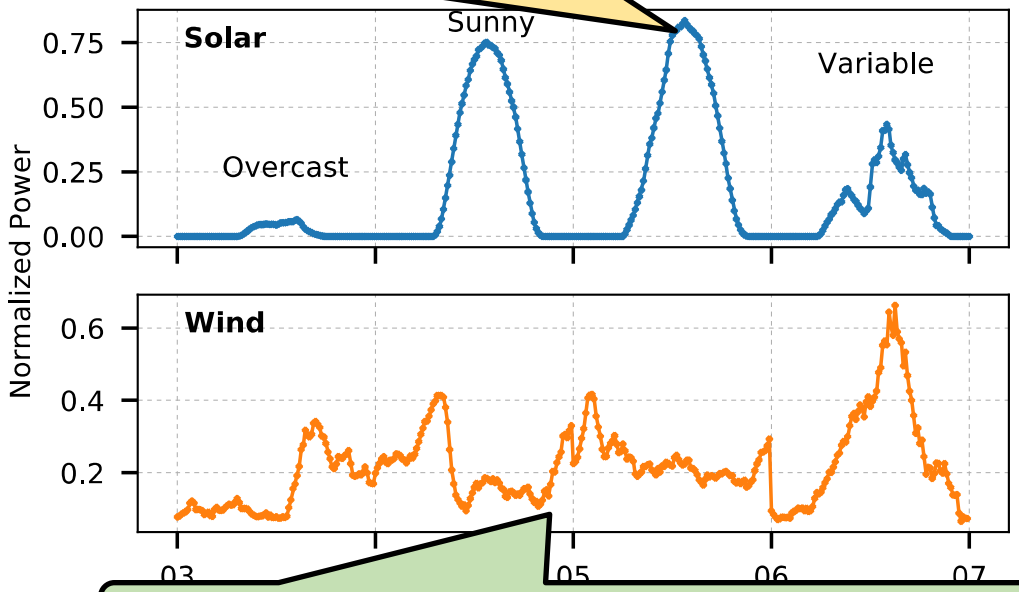


Renewable energy is effective to reduce the carbon footprint

Deploying Renewable-Based Modular Data Center Is Challenging

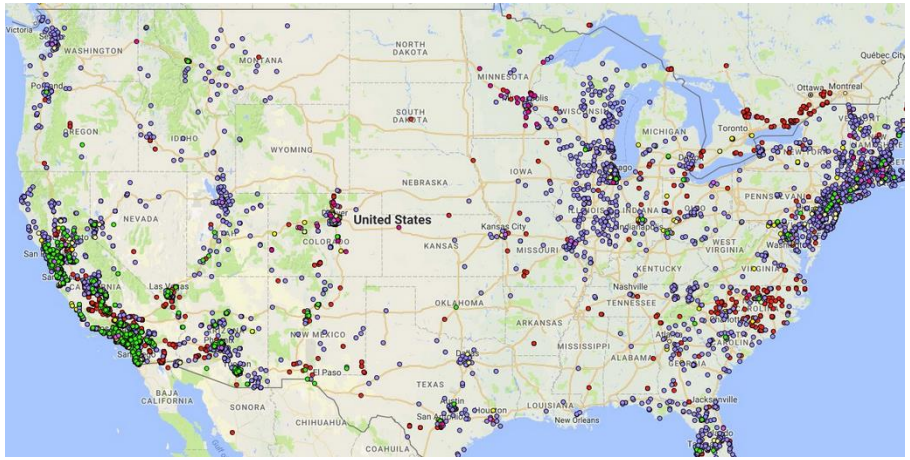
Temporal Variation

Solar power follows the diurnal cycles



Wind power depends on the weather condition

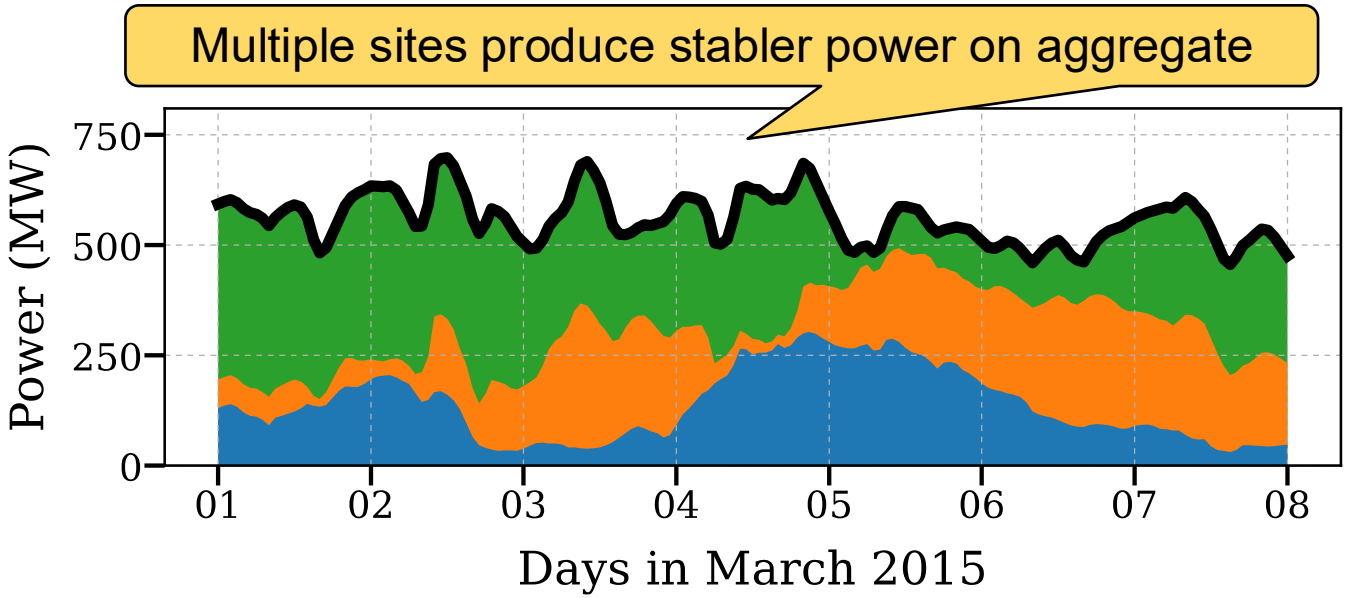
Regional Variation



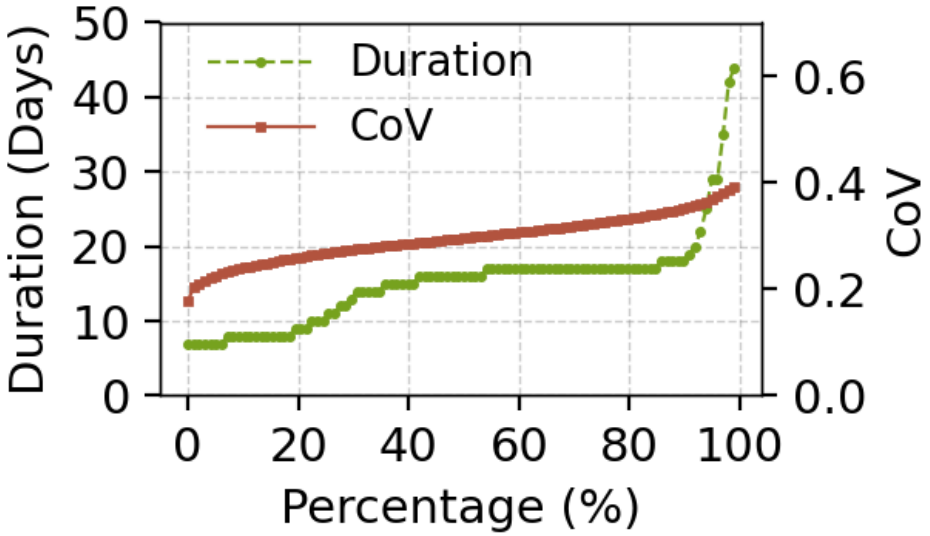
Thousands of different renewable farms across the U.S. and Europe

How to deploy data centers to efficiently use renewable energy ?

Insight 1: Complementary Production Patterns across Renewable Farms



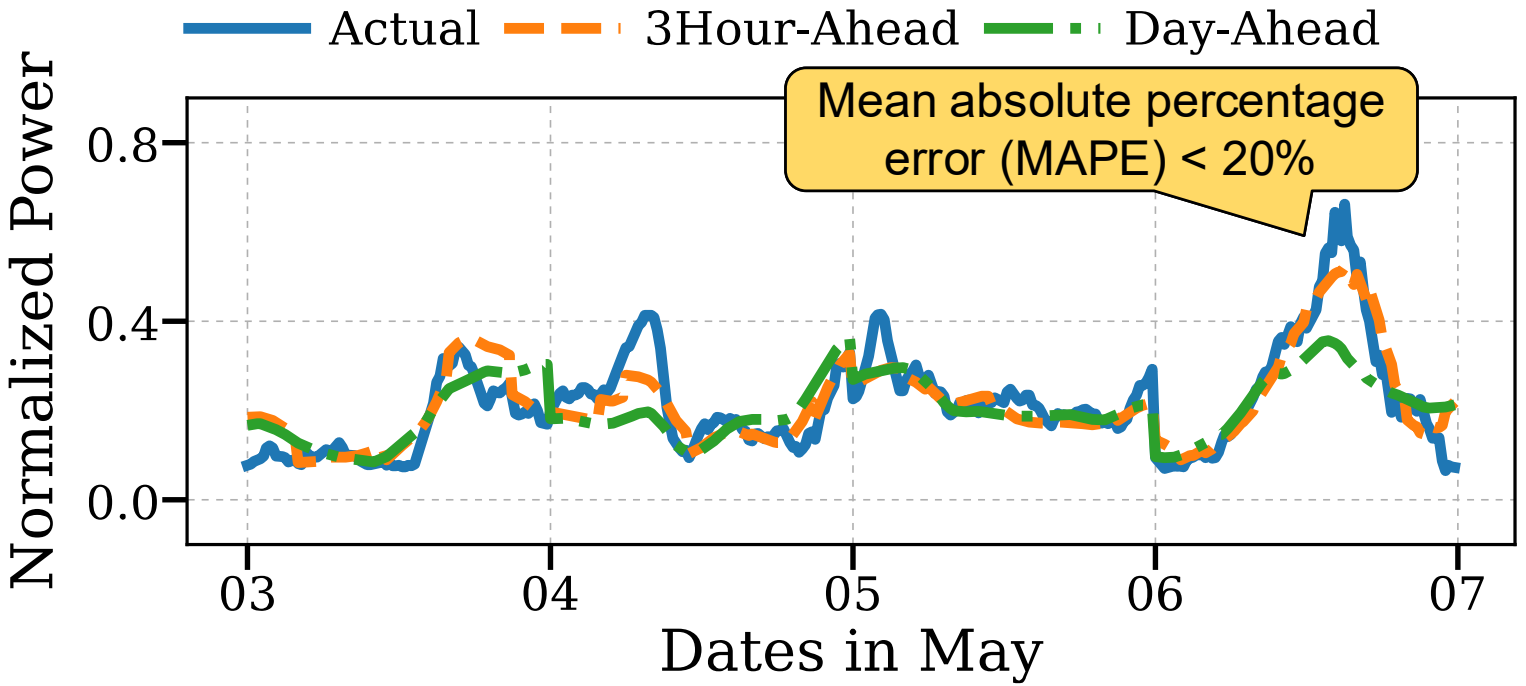
There are complementary production patterns across multiple renewable sites



Complementary sites preserve for a sufficiently long time

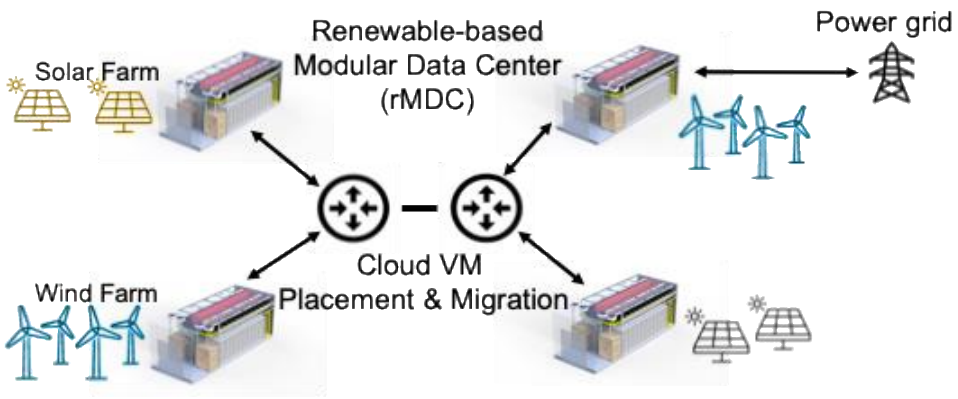
We can leverage the **complementary** patterns of renewable sites to provide stable energy source!

Insight 2: Good Predictability of Renewable Energy Production



High prediction accuracy of hours-ahead renewable energy production

SkyBox: Exploring the Efficiency of Renewable-Based Modular Data Centers

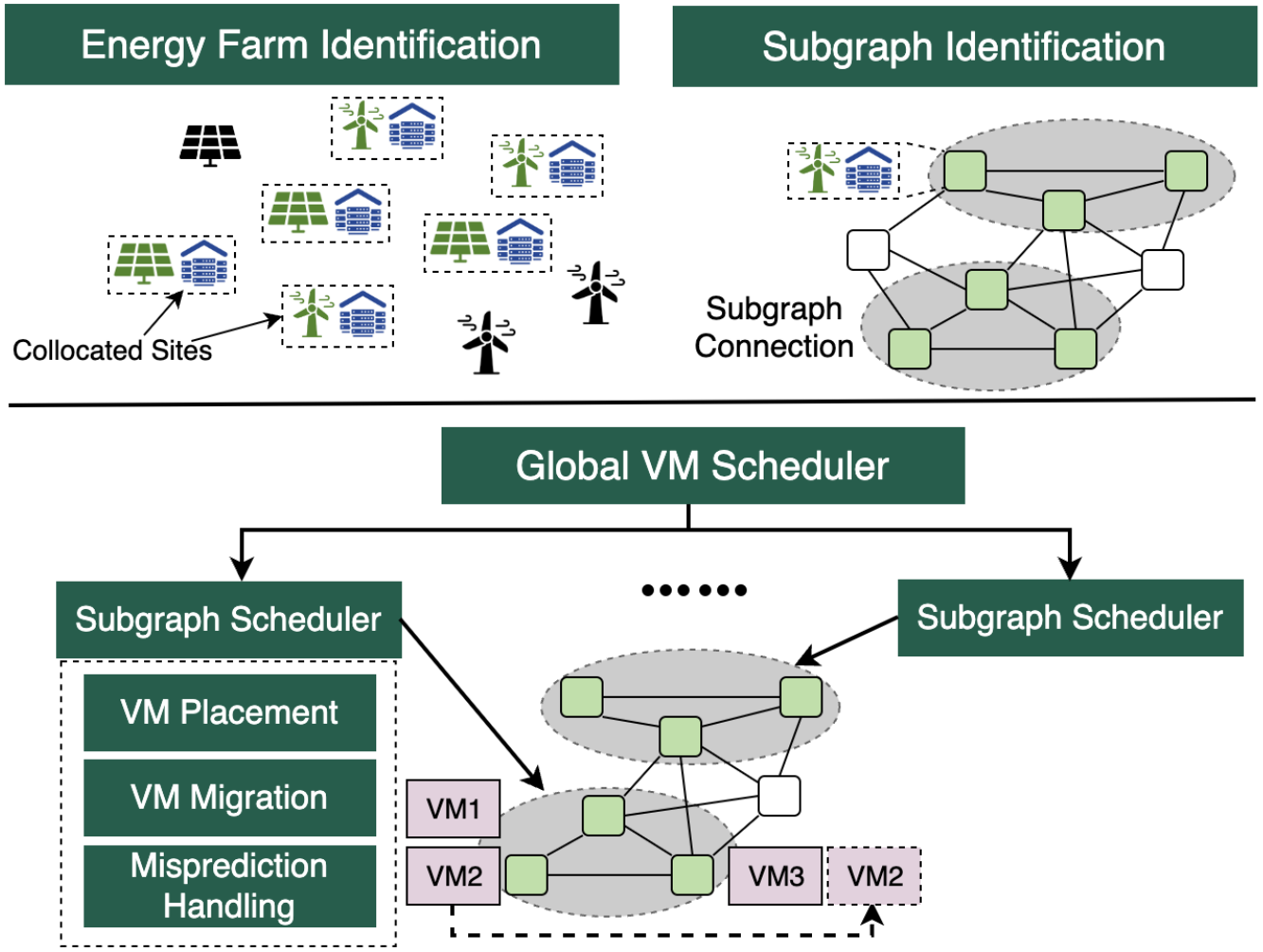


The target system architecture of SkyBox

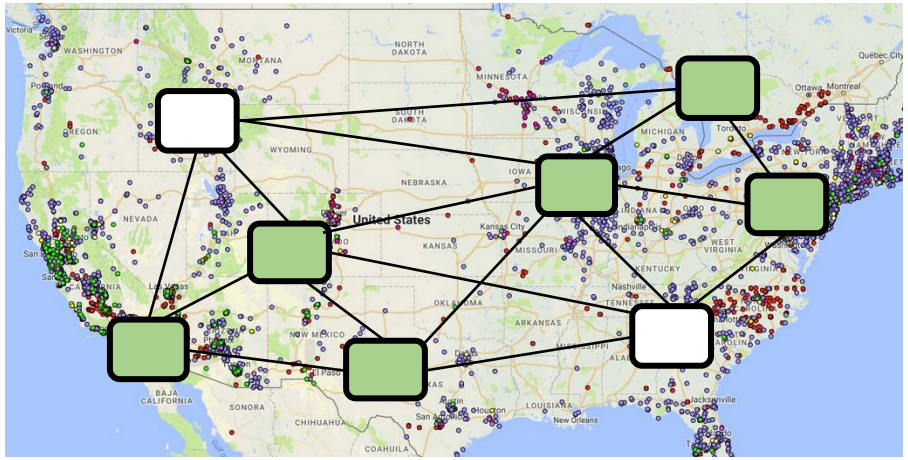
How to Select Renewable Farms?

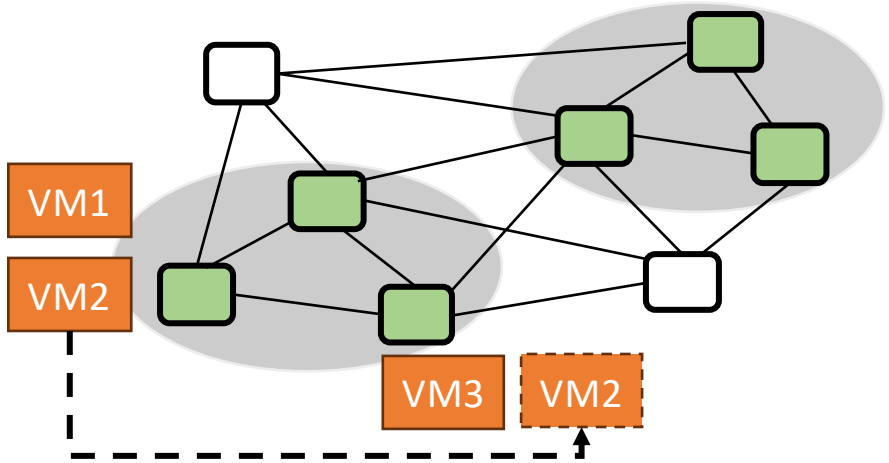
How to Identify Complementary Sites?

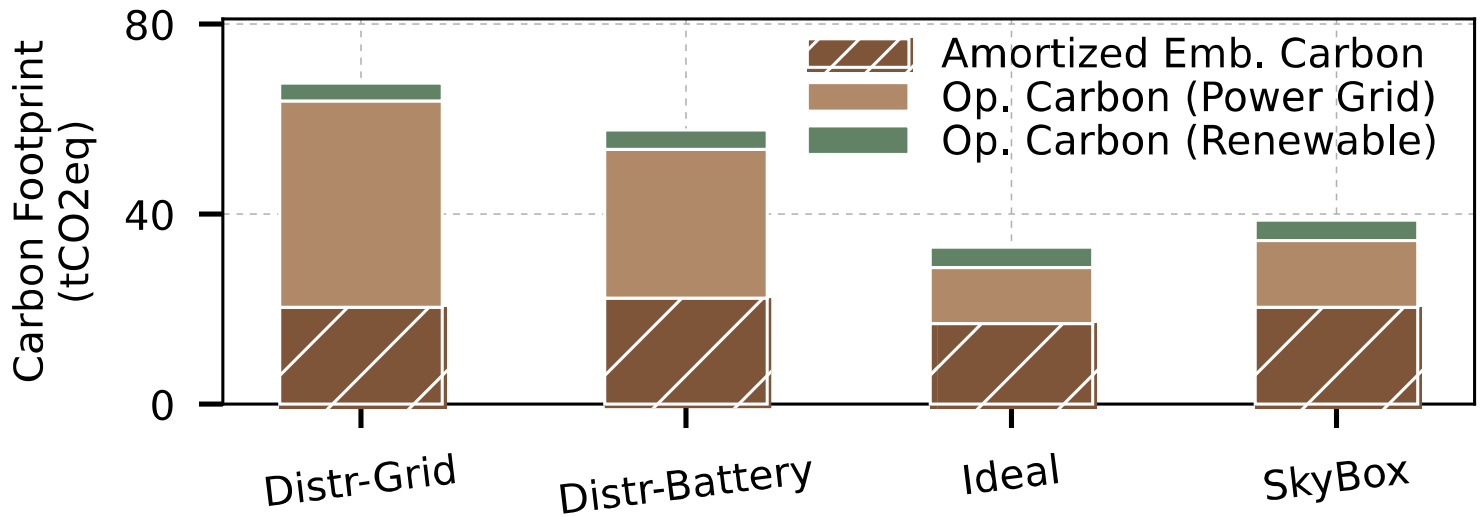
How to Schedule VM Workloads?



Benefits of SkyBox on Reduced Total Carbon Footprint

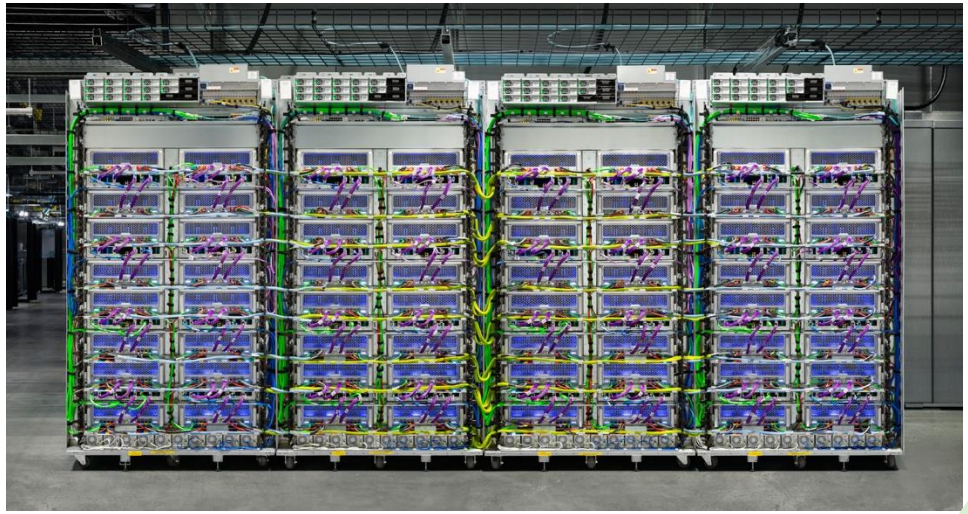


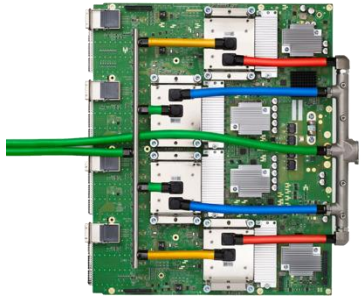




SkyBox can effectively reduce the total carbon footprint by 39% (up to 46%)

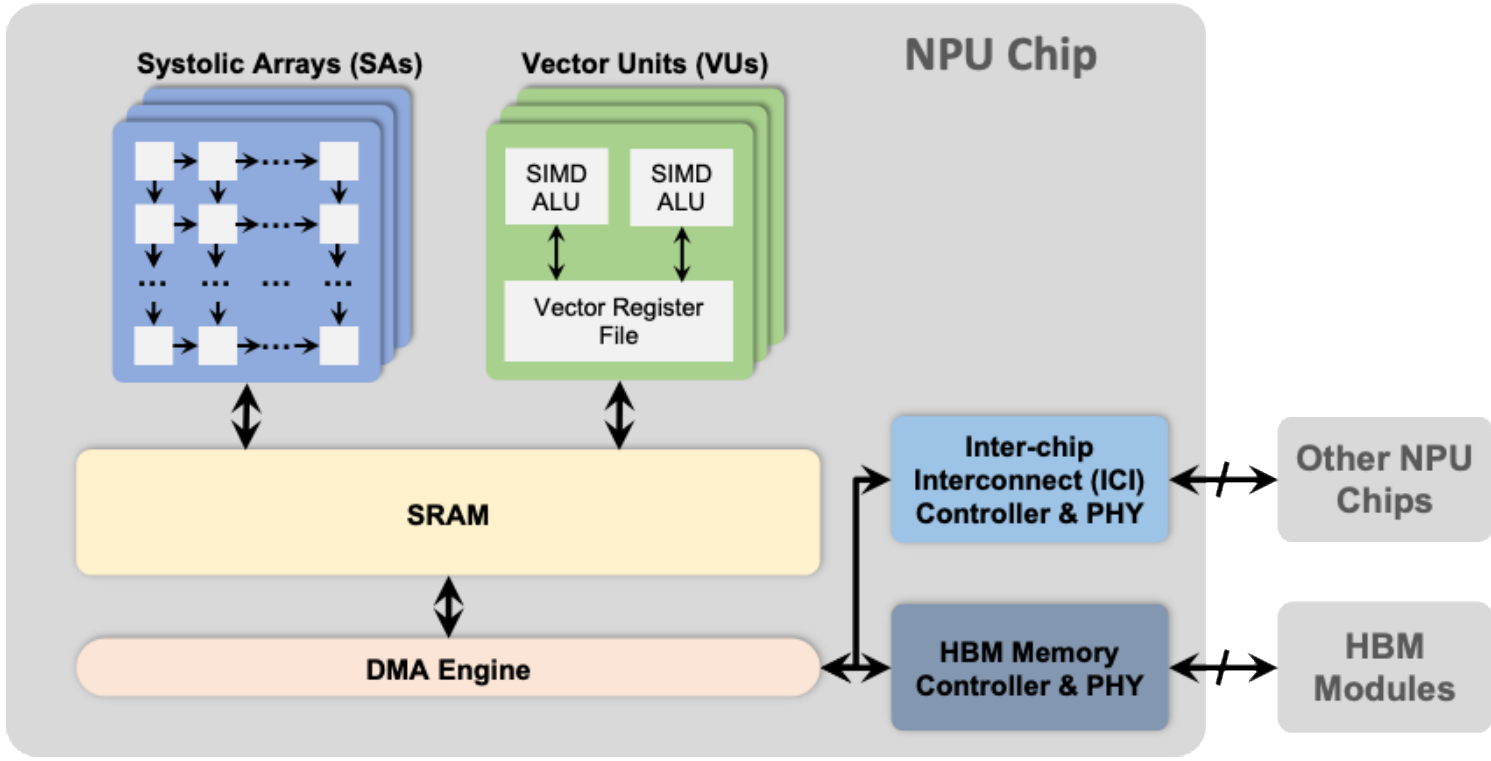
Energy-Efficient AI Chip Design and Implementation



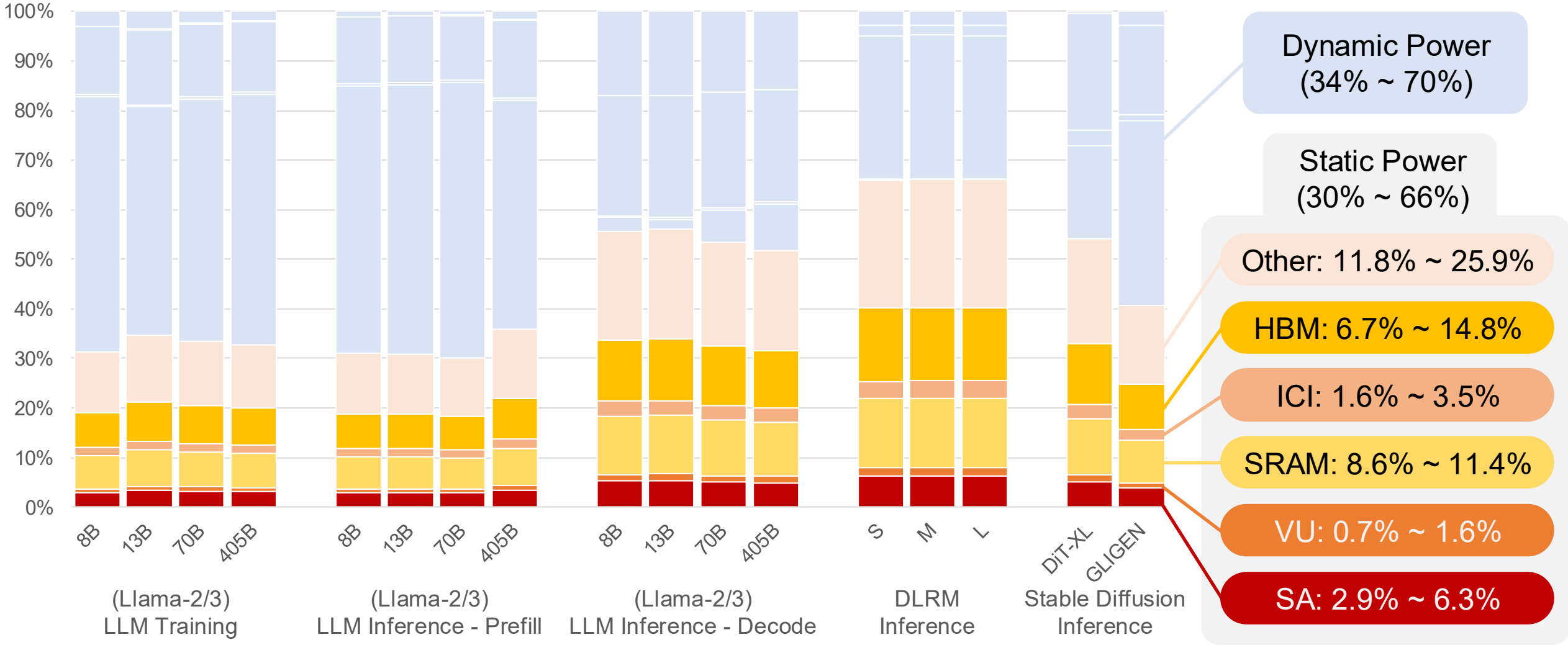


Google TPU

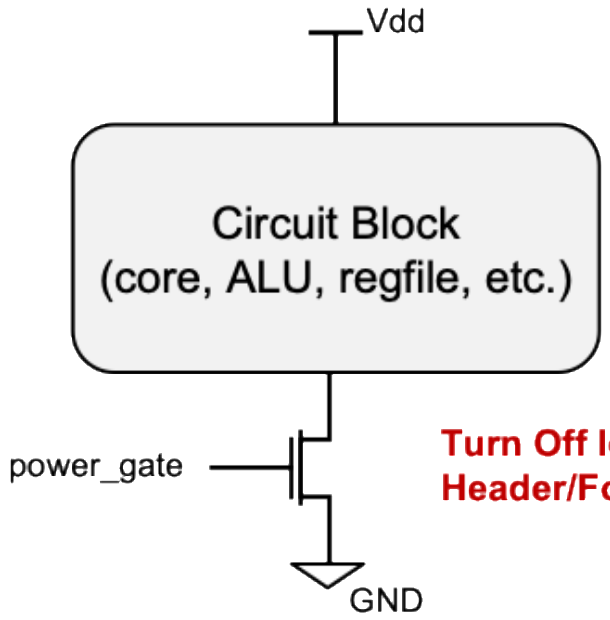
Microarchitecture is also critical !



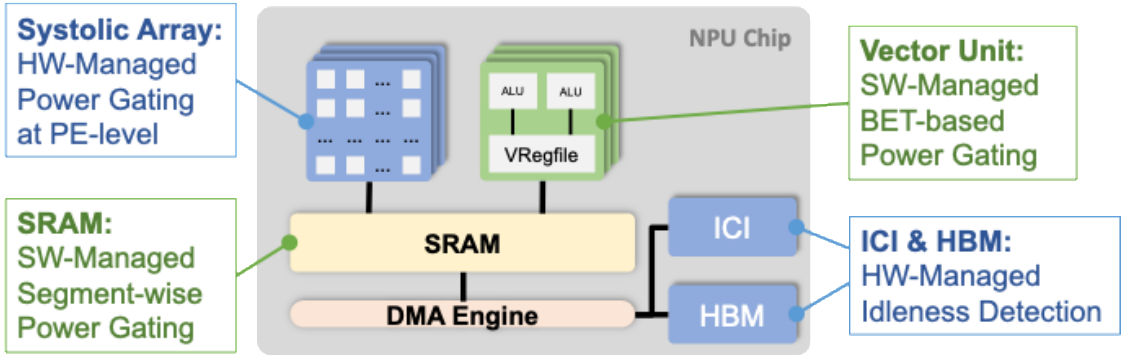
NPU Chips Today Suffer From Significant Energy Waste



ReGate: A Holistic NPU Power Gating Solution with HW/SW Co-design



Turn Off Idle Components with Header/Footer Transistor



Architectural Support for Fine-Grained NPU Power Gating

NPU ISA Extension for Software-Driven Power Management

Compiler-Driven Precise NPU Power Gating

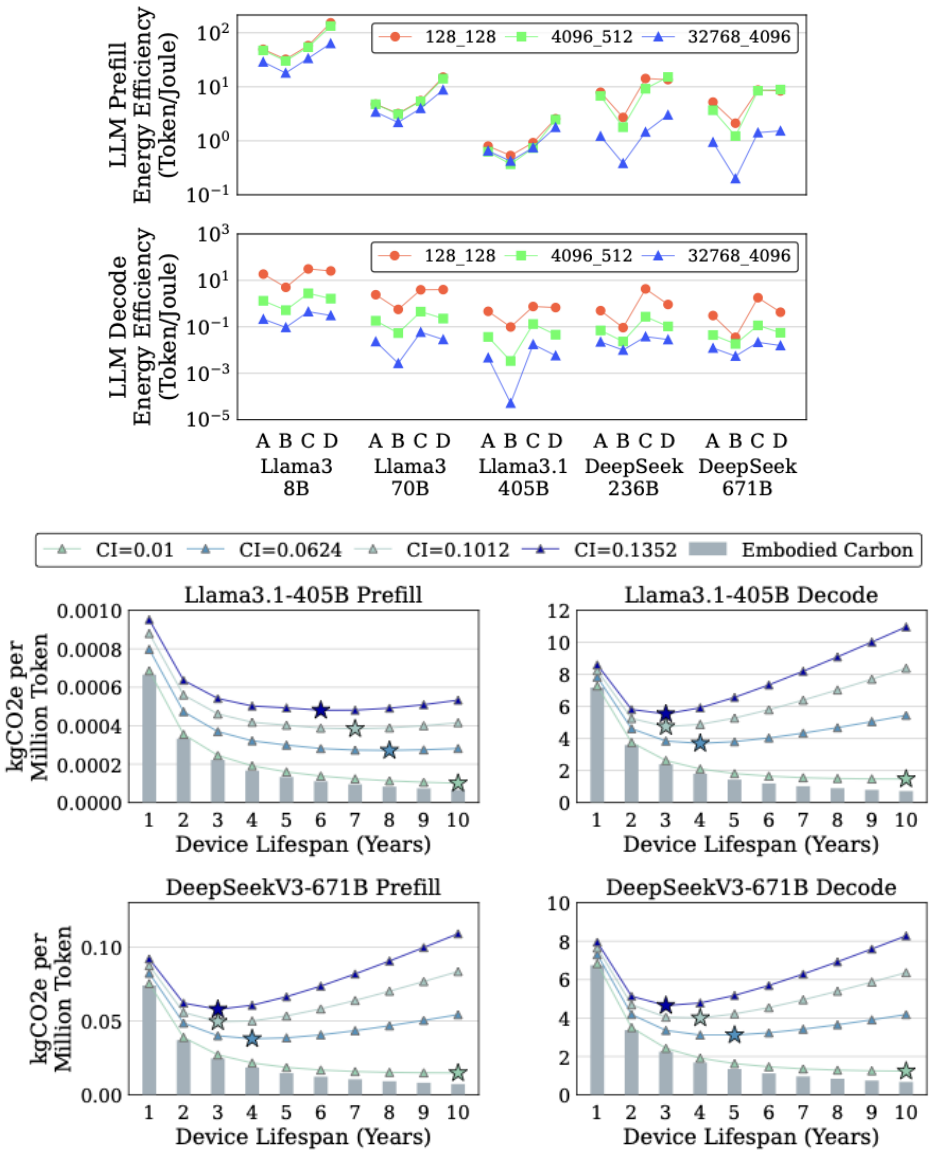
Up to 32.8% Energy Saving with Negligible Performance Overhead

Reuse Different Generations of AI Chips for Cost-Effective LLM Serving

TABLE I: Specifications of different NPU chip versions. The data for NPU-A/B/C/D are derived from TPUv4/5e/5p/6e. "*" means the parameter is not officially disclosed, and it is inferred from public data and our experiments.

	NPU-A	NPU-B	NPU-C	NPU-D
Release Year	2021	2023	2023	2024
Process Node	7nm	7nm*	7nm*	4nm*
TFLOPS (bf16)	275	197	459	918
Max. TFLOP/Joule (bf16)	1.07*	0.72*	1.16*	3.50*
Systolic Array Width	128	128	128	256
SRAM Size (MB)	128	64*	128*	128*
HBM Bandwidth (GB/s)	1200	819	2765	1640
HBM Size (GB)	32	16	95	32
HBM Type	HBM2	HBM2E	HBM2E	HBM2E*
Max. HBM GB/Joule	11.06*	6.94*	18.51*	13.52*
ICI BW/link (GB/s)	50	50	100	112
# of ICI links/chip	6	4	6	4
ICI Torus Topology	3D	2D	3D	2D
Max. # of chips/pod	4096	256	6144	256

AI chips have been developed and evolved at fast pace



Takeaways of My Talk

- We need a holistic approach to developing sustainable AI infrastructures.
- We need renewable energy to achieve the net-zero goal. However, the deployment of datacenters is challenging.
- Energy-aware computing system support is necessary.
- Great opportunities in energy-efficient AI chip design and implementation.
- Reusing hardware components is challenging but rewarding (e.g., cost and carbon reduction, energy efficiency).
- Many exciting topics, look forward to more collaborations.

Thank You!

Jian Huang

jianh@illinois.edu

Systems Platform Research Group

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN