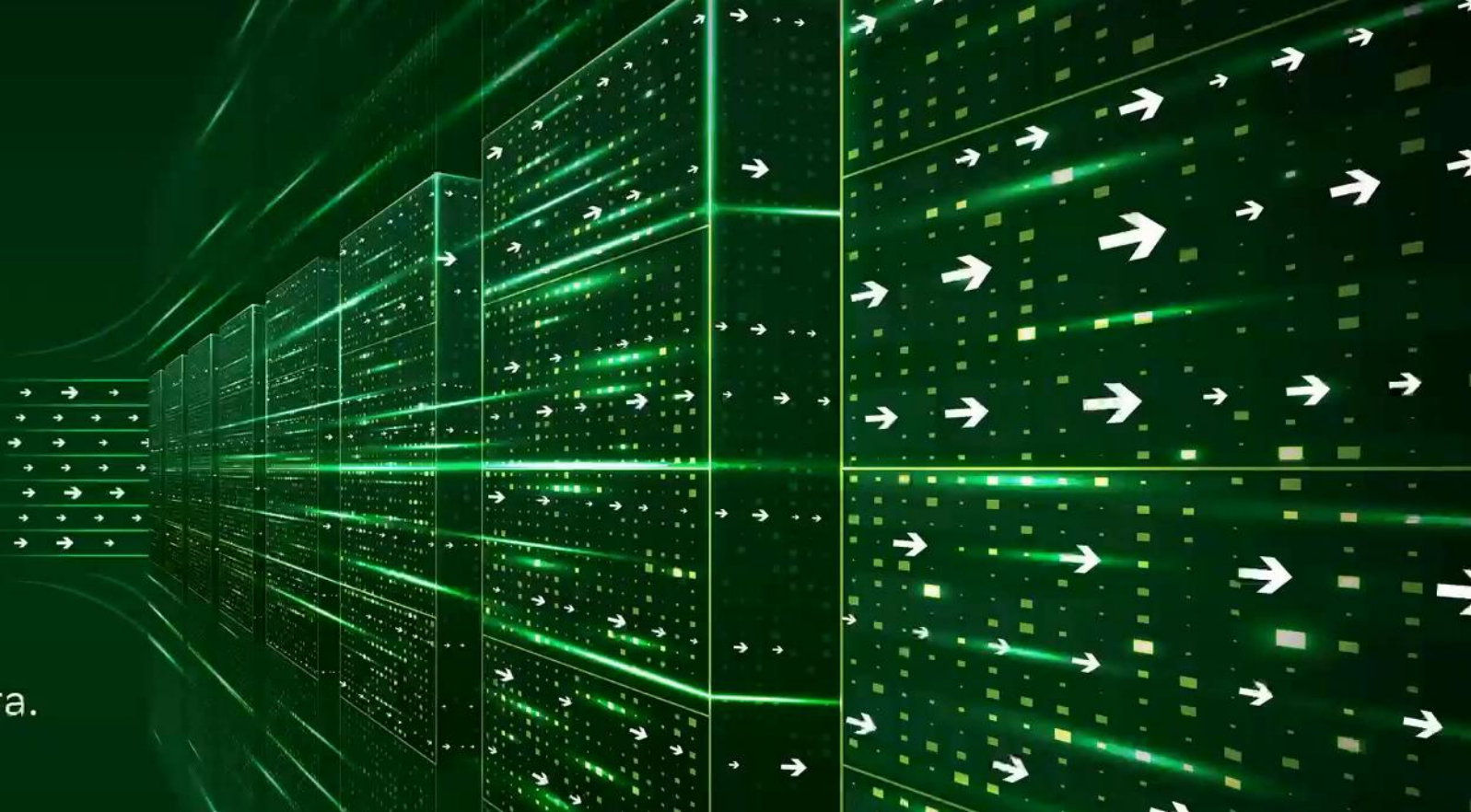




a.



OCP
GLOBAL
SUMMIT

Scaling Innovation for the AI Era.

October 12-15, 2026 | San Jose, California

Powering the AI Era

Building sustainable and efficient infrastructure at cloud scale



Brijesh Warriar

Vice President, Software Engineering



Nithish Mahalingam

Partner Director, Software Engineering

AI is driving demand faster than we can build

Power, servers, and capital are all scaling at once

AI is driving demand faster than we can build

Power, servers, and capital are all scaling at once

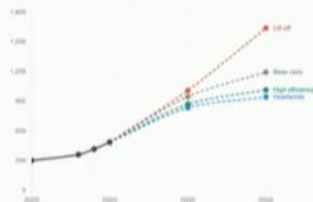
POWER

2.5x

Datacenter electricity demand by 2035: > 137 GW base case, > 188 GW with accelerated AI (IEA)

IEA scenarios of electricity demand from data centres

Global data centre electricity consumption (TWh). Four IEA scenarios for 2020 and 2035.



Note: 2020 figures are preliminary given the high level of uncertainty around data centre demand growth. Data source: IEA, Key questions on Energy and AI (2024).

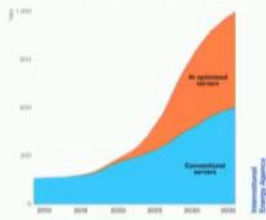
SERVERS

+31%

AI server shipments in 2026 YoY, vs +12.8% for all servers

Data centre electricity demand is set to surge in the next decade, driven by AI

Data centre electricity demand, historical & projected through 2035



CAPEX

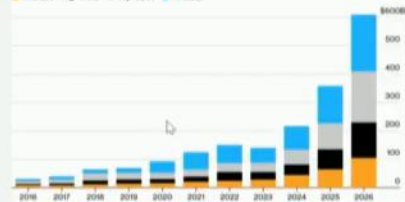
~\$650B

Hyperscaler capex in 2026 on servers and datacenters (Bloomberg)

Hyperscaler Capex Explodes Higher

2026 estimates based on mid-point of guidance (META, GOOGL, AMZN), Bloomberg consensus (MSFT)

Microsoft Meta Alphabet Amazon



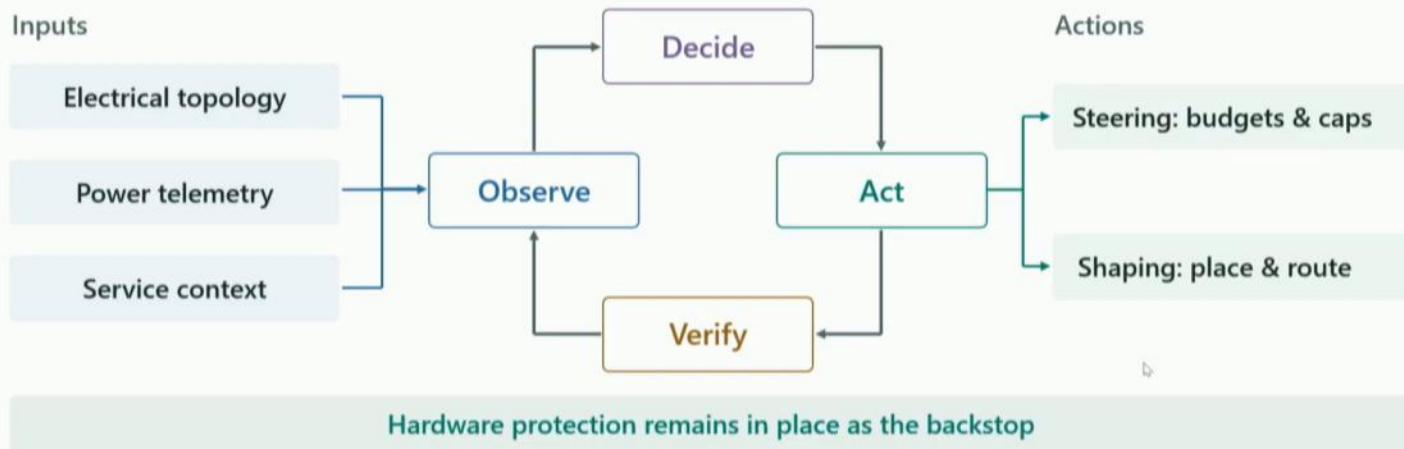
Source: Bloomberg

Note: Capex based on fiscal year-ends, Dec. (AMZN, GOOGL, META), June (MSFT)

Bloomberg

Power is a cross-stack problem

Software-defined power

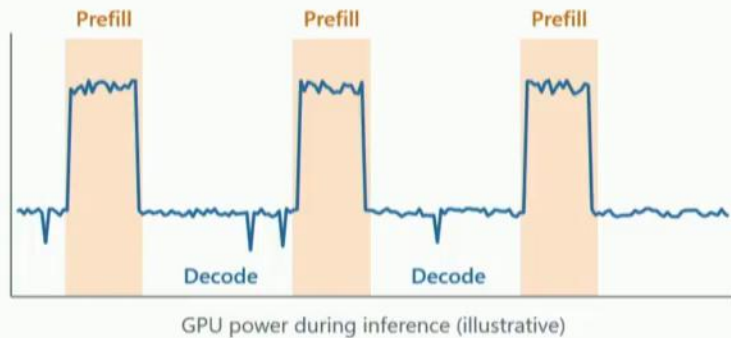


Measure power, set policy, and control it in software

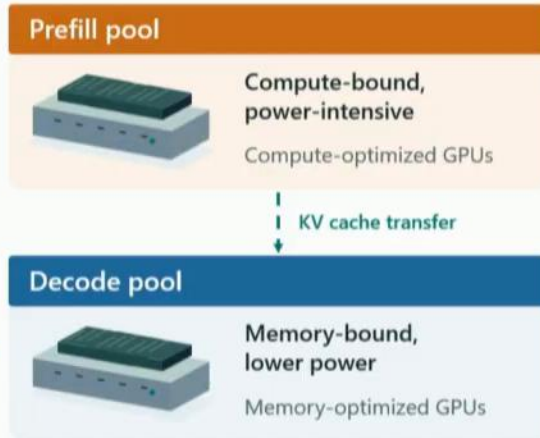
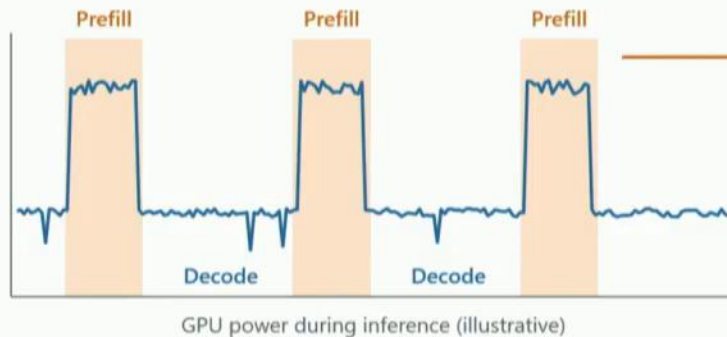
Power steering: hierarchical power management



Inference: splitting the phases



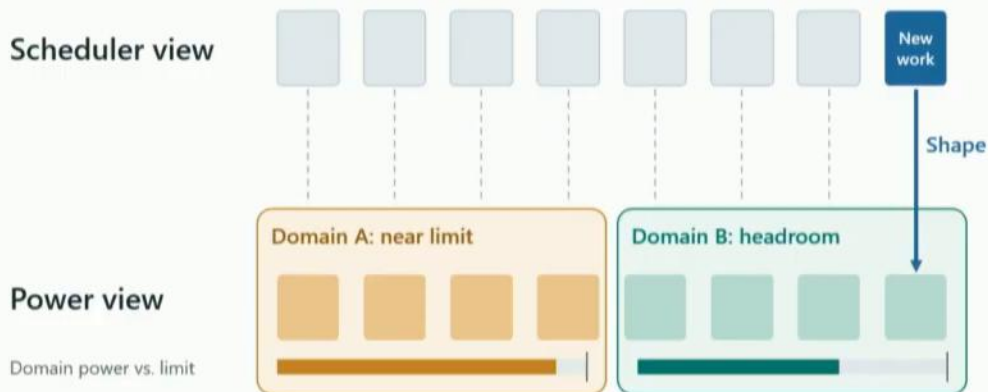
Inference: splitting the phases



Power shaping: placing work where the power is

Power shaping: placing work where the power is

Schedulers see identical nodes; power shaping uses the power view to place work



Shape the work to the power: new work goes where the domain has headroom

Power-aware inference



Incoming requests



Domain A: near limit



Domain B: headroom

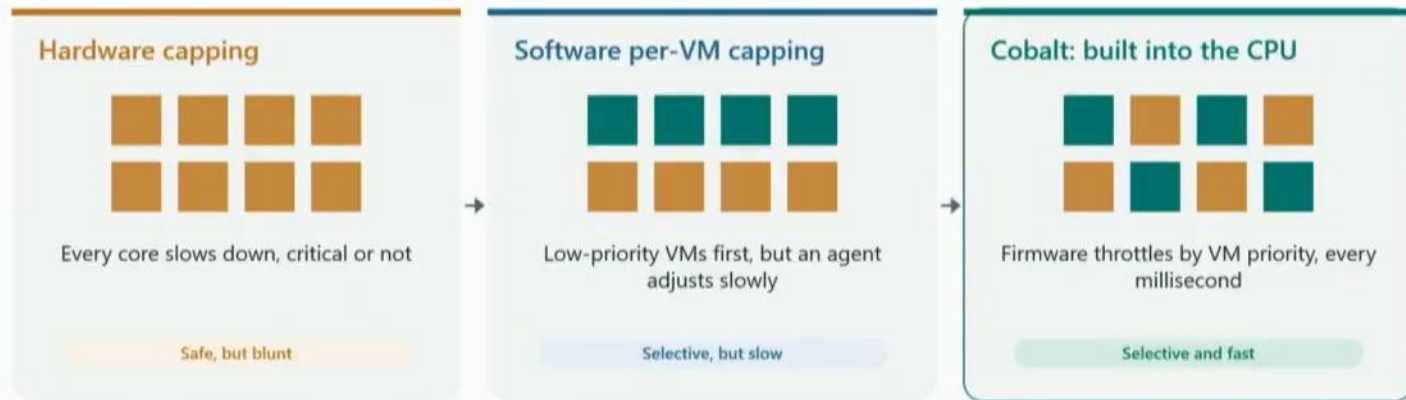


Workload-aware power operations

Per-VM power capping in silicon: Azure Cobalt



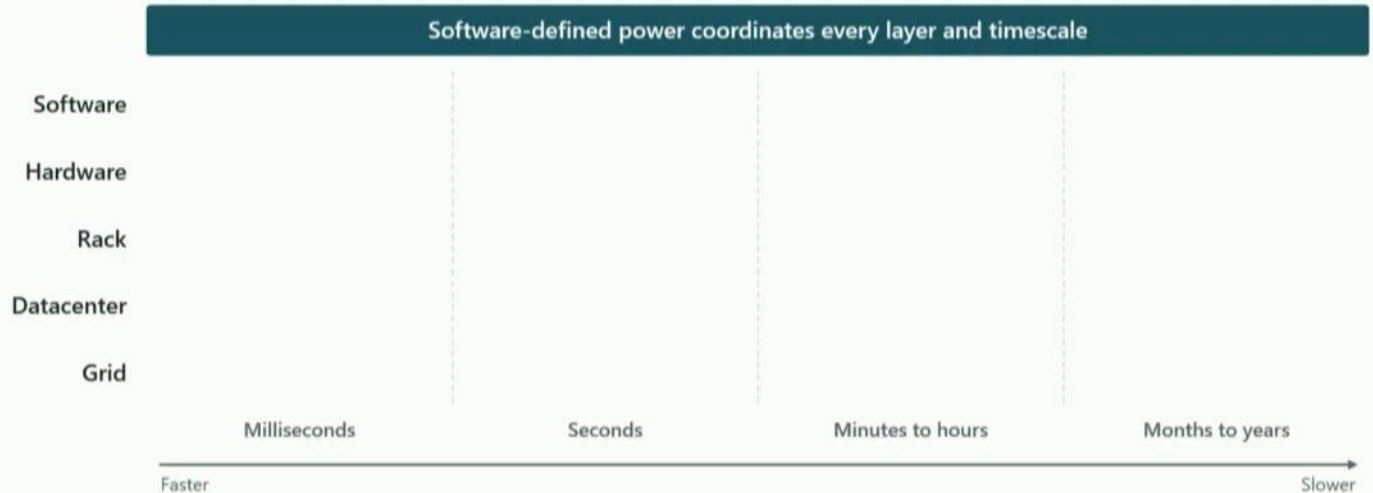
The CPU throttles by VM priority, so critical work keeps running when power is tight



Up to 20% more power oversubscription on Cobalt 200, with critical VMs protected

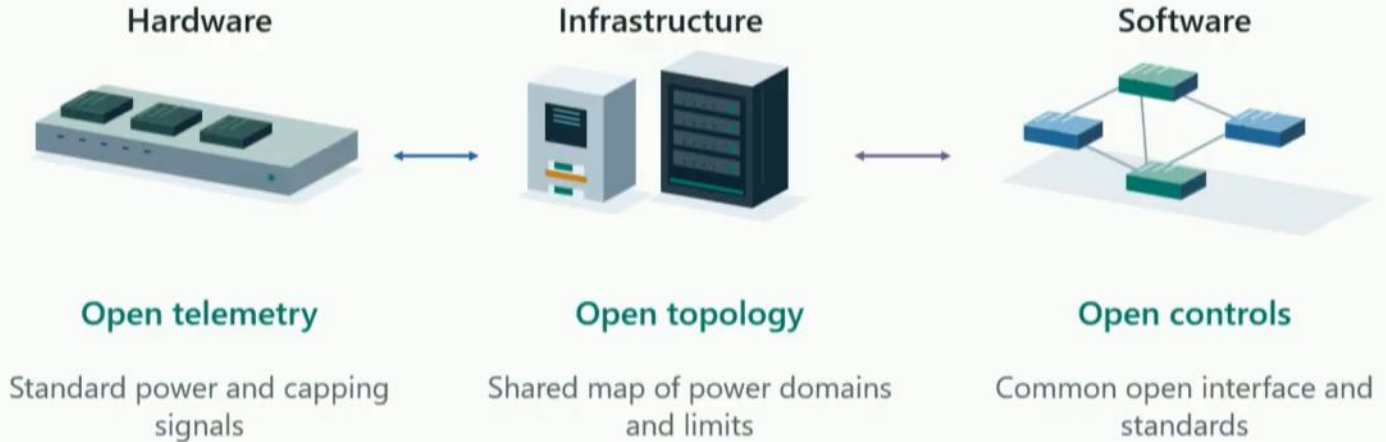
Putting it all together

Each technique acts at a different layer and timescale; together they cover the stack



Where OCP can help

Open, standard interfaces so every operator can make power software-defined



Thank you!