Red Hat Summit

Connect

Sustainable Computing for the Future

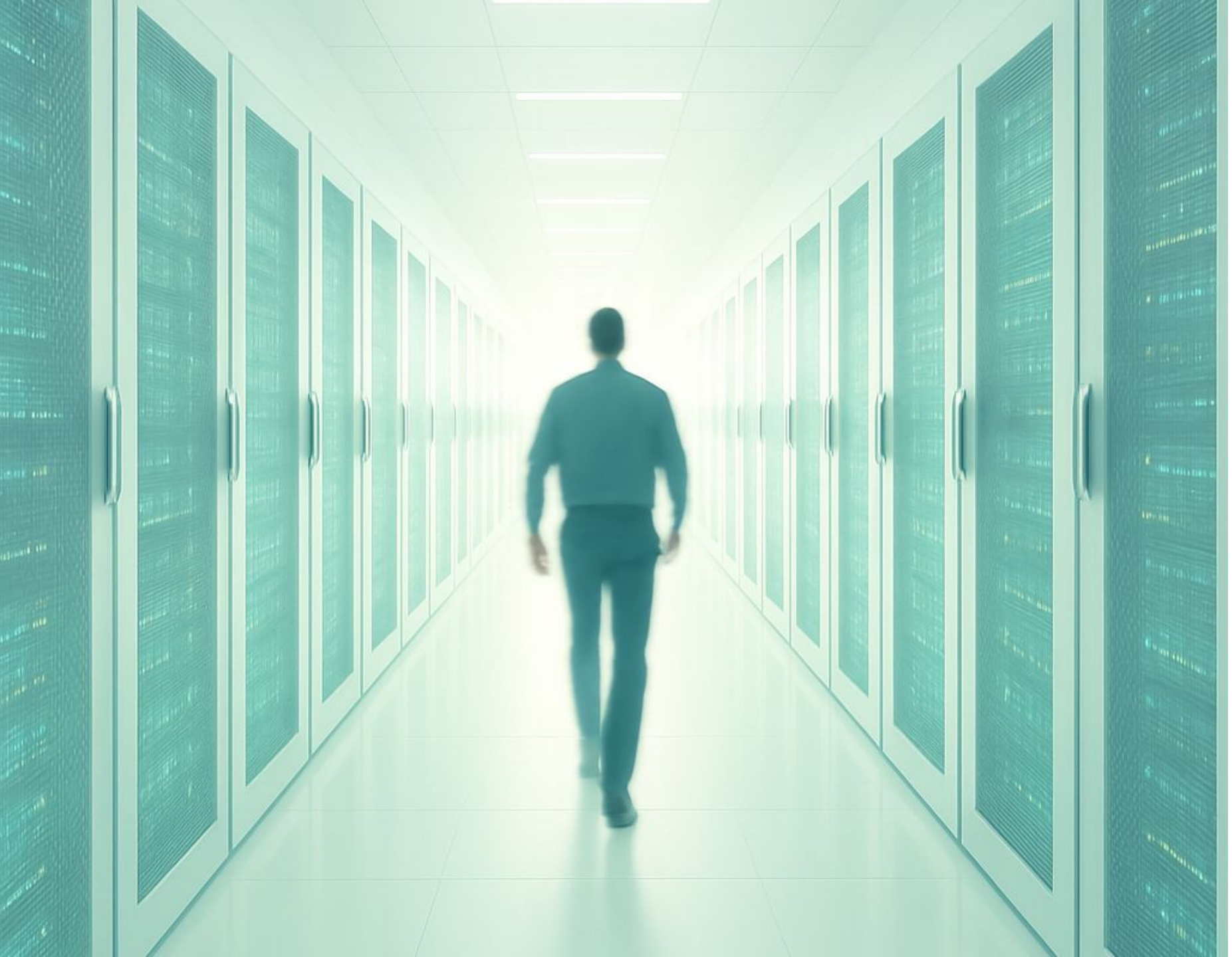
Turning Constraints into Resilience in the AI Era

Waqaas Kausar

Principal Solution Architect, UK Public Sector

What will be talking about?

- The surge in compute demand
- Emerging constraints across the system
  (why it's different this time)
- Systematic actions we can take in applications & systems
- How Red Hat enables sustainable platforms



The Surge in Compute Demand

“**Compute power is emerging as one of this decade’s most critical resources.**”

McKinsey, Apr 2025 : “The cost of compute: A $7 trillion race to scale data centers”

Why Compute Matters More Than Ever

Compute power is becoming the new critical resource.

Business continuity, AI adoption, and digital growth hinge on access to compute.

Compute is no longer just a cost driver, it is a strategic asset.

The New Scale of Demand

AI workloads are multiplying demand far faster than historical growth.

Training + inference both require major new capacity.

Unlike in the past, demand is global, constant, and system-wide.

Emerging Constraints Across the System

“GenAI is creating an insatiable demand for power that will exceed the ability of utility providers to expand their capacity fast enough”

Gartner, Nov 2024: “Power Shortages Will Restrict 40% of AI Data Centers By 2027”

Multiple Constraints, Not Just Data Centre Capacity

System-wide pressures are reshaping computing

- Energy & Carbon – supply limits, sustainability pressures.
- Power Grids – connection delays and reliability risks.
- Water & Land – cooling demand, siting conflicts, community impact.
- Supply Chains & Skills – equipment, components, talent shortages.
- Regulation & Finance – ESG, disclosure, and investor scrutiny.

Source:
[1] International Energy Agency - AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works

Infrastructure Constraints: Power Demands

AI workloads are reshaping the global energy map

- Data-centre power demand is currently ≈ 415 TWh and projected to double to 945TWh by 2030, close to Japan's entire demand today. [1]
- AI-optimised data-centre energy use was 23 TWh in 2022 and is set to grow 44.7 % CAGR to 146 TWh by 2027 - more than Sweden's 2021 electricity demand. [2]
- GPT-3 training consumed ~1,287 MWh of electricity, emitting 502 metric tons of CO_2, orders of magnitude above classic apps [3]
- Inference will become an area with huge scope for efficiency gains.[4][5]

Source:
[1] International Energy Agency - AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works
[2] IDC - Sustainable AI for Sustainability
[3] Wharton Business School, UPenn - The Hidden Cost of AI Energy Consumption
[4] WEF, Apr 2024 - How to manage AI's energy demand — today, tomorrow and in the future
[5] Columbia Climate School, June 2023 - AI's Growing Carbon Footprint

Infrastructure Constraints: Power and Grid Limitations

- Ageing power grids
  - **UK** – more than 1700 energy generation projects in decade long queue [1]
  - **Europe** – 1.7 TW of renewable schemes are stuck across sixteen countries [2]
  - **US** – 2.6 TW of solar, wind and storage waiting for connection [3]
- North American Electric Reliability Corporation (NERC) warns AI demand and data centre growth threatens reliability [4]
- Data centres are being held back from connection to the grid
  - Ireland – No new data centres in Dublin, others will have to create their own generation and storage until 2028
  - Netherlands – no more grid capacity for data centres in Amsterdam city limits
  - Germany, new DC connection as far off as 2031; UK connection queue as long as 13 years [5]

[1] The Guardian, Jan 2025 - Great Britain's energy grid operator closes connection queue
[2] Beyond Fossil Fuels, May 2025 - Outdated grid planning and weak governance stalling Europe's transition away from fossil fuels
[3] Energy Markets and Policies, Berkley Lab, Jan 2025 - Grid Connection Barriers To New-Build Power Plants In the United States
[4] FT, Apr 2025 - Is the US power grid ready to meet the demands of data centres?
[5] Aggreko, Oct 2022 - How European data centres can thrive in the era of grid uncertainty

Infrastructure Constraints: Water, Land, and Zoning

- A typical hyperscale data centre can consume 3-5 million gallons (11 to 20 million litres) of water per day. Enough for a town of 30 - 40,000 people [1]
- U.S. data centres used an estimated 626 billion litres (165 billion gallons) of water in 2020, primarily for cooling. [2]
- Water-intensive cooling facing backlash in dry regions
- Land scarcity near population centers increases costs
- Zoning rules restrict noise, emissions, and visual impact

[1] Data Centre Magazine, Nov 2023 - Transforming data centre cooling for a sustainable future
[2] Nature, Feb 2021 - Data centre water consumption

Infrastructure Constraints: Supply Chains and Skills

- Delays in chips, HVAC, transformers, generators
- Scarcity of facility engineers and automation experts
- Talent and equipment shortages slow operational readiness

Why it’s different this time

In the past: More servers = more compute.

Today: Constraints interact.
Energy affects carbon, grids, water, land, and financing.

The system is under simultaneous pressure, which changes platform design.

What can we do about this?

“The growing demands of these complex models are raising concerns about AI’s environmental impact.”

HBR Jul 2024 :
“The Uneven Distribution of AI’s Environmental Impacts”

Applications: Smarter by Design

Designing workloads for maximum efficiency

- Efficiency first: more useful work per watt.
- Smaller models, distributed inference, privacy-friendly deployment.
- Continuous optimisation of workloads.

Systems: Resilient Platforms

Building flexible, energy-aware infrastructure

- Hybrid distribution across cloud, edge, and on-prem.
- Heterogeneous hardware fit to workload.
- Carbon- and energy-aware scheduling.
- Observability to track energy + carbon footprint.

Operations: Automating Sustainability

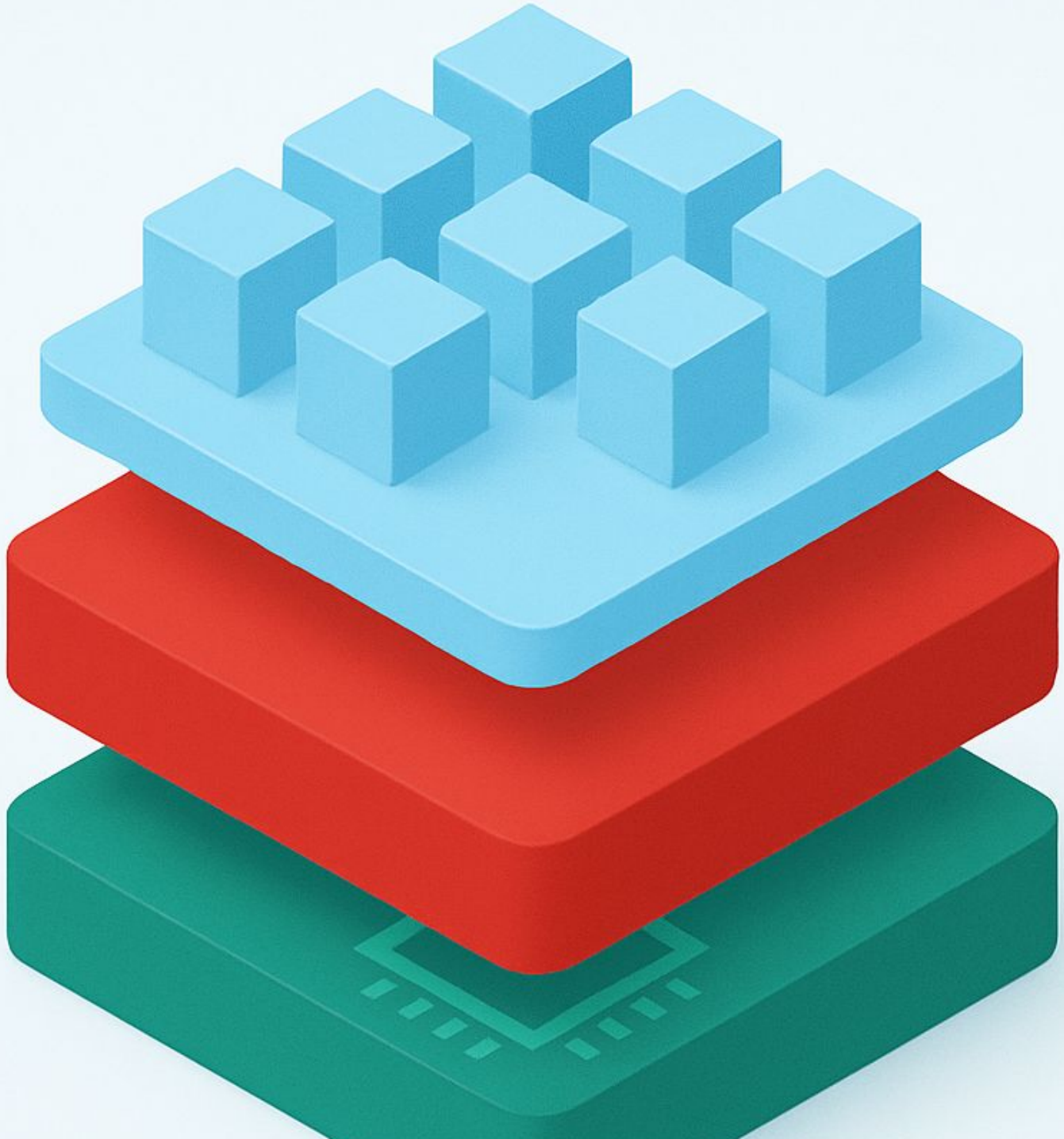
Making sustainability a built-in operating principle

- Policy-driven automation to enforce energy and ESG rules.
- Demand-response readiness to align with grid availability.
- Circularity: extending hardware life, recycling rare earths.

Sustainable ≠ Slower

Rethinking Efficiency

- Smart placement beats brute force.
- Automation and policy can reduce energy without reducing impact.
- Optimised platforms outperform legacy even at lower consumption.



How Red Hat Enables Sustainable Platforms

“We’re going to start seeing smaller models … For many customers it’s crucial, from an efficiency and data-privacy point of view, that these models can run on a very small form-factor device—even a CPU.”

Erin Boyd, Distinguished Engineer, Red Hat

What Resilient Platforms Require

Characteristics of Sustainable Platforms

1. Energy-aware automation for open, cloud-native workloads for portability
2. Hybrid workload distribution (cloud, edge, on-prem)
3. Heterogeneous hardware (x86 + Arm + GPU + DPU) per job
4. System-wide observability of energy & carbon
5. Community innovation via open source

Red Hat's Role

Enabling Sustainable Architectures

- **Red Hat OpenShift**: Carbon-aware scheduling and portability, autoscaling, Kube-native tools.
- **RHEL + Edge**: Lightweight, update-efficient, reduced data centre dependence.
- **Ansible Automation**: Codify energy and ESG policies into workflows.
- **Red Hat AI**: Focus on efficiency
  - Train and fine tune smaller models
  - Scale and deploy distributed inferencing

Case Study – Valencia City Council

- Red Hat OpenShift enables Valencia City Council to free up more than 4,000 hours of work, increase efficiency and reduce carbon footprint
- 5 Mt CO_2e avoided via Red Hat Open Hybrid Cloud
- OpenShift + Ansible cut energy cost 30 % while improving uptime
- Business, environmental & resilience wins

Source:
Red Hat - Valencia City Council Pioneers e-Government with Red Hat

Wrap up

Key Takeaways

- Compute demand is growing massively and rapidly.
- Sustainable compute will be a business-critical metric.
- AI accelerates urgency - grid & capacity constraints are real.
- Open source platforms make sustainable scale possible today.
- Red Hat + ecosystem can help you act now.

Red Hat Summit

Connect

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat