



HYPERION RESEARCH

Considerations for Sustainability

ISC24 Market Update Briefing
May 2024

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

Not Your Father's HPC... ..Now Add AI

An Almost Untenable Increase in Complexity

HPC + AI

Research
Engineering
Scientific
Discovery
Business
Analytics
LLMs
SLMs
Multimodal
GenAI



Prioritizing Sustainability in Procurements

Sustainability and energy efficiency considerations will become a dominant factor in many procurements

- **New CPUs/GPUs/xPUs have substantially larger power requirements**
 - Maximize performance capabilities for these compute elements, consistent and affordable energy
- **Energy costs**
 - Energy crisis / inflation / geographic concern
- **Energy consumption legislation**
 - Recent legislation on the energy consumption of data centers
 - Implementation: energy-conscious procurements, renewable energy where possible, and optimizing workloads for efficiency
- **Environmental stewardship**
 - Perception of environmental stewardship amongst HPC users
 - Metrics and requirements to identify sustainable vendors
 - Forward-thinking vendors want to respond to the call
- **New technologies are needed to improve energy usage**

Energy Costs to Surpass Maximum Performance in System Tradeoffs

The cost of energy will more frequently outweigh HPC performance increases as some sites will settle for “good enough” speeds in order to reach energy efficiency goals

- **Many HPC sites already find themselves constrained by the escalating costs of energy consumption due to increasing power requirements of highest performing computational processors**
 - Large AI training models are expected to grow at a rate that will require major increases in power
- **Attention will turn towards a pragmatic tradeoff and balance between affordability and performance**
- **Focus on optimizing performance per unit of energy consumed particularly acute in regions of high energy costs, such as in Europe and Asia**

A Datacenter Perspective

Multiple challenges magnified by adoption of AI

- **Power demands and constraints**
 - From 10s of MWs → > 1GW
 - Generation, transmission, or distribution?
 - BYOP (bring your own power)
 - 10s of racks being replaced by a handful of racks for equivalent performance
- **Rack density – from 10KW → 40KW → 400KW**
- **Retrofit or new build for data center expansion?**
 - Access to adequate power for current and planned needs
 - Ability of existing facilities to support internal distribution of more power
 - Readiness for and ability to support current and planned liquid cooling requirements
 - “Time to power” for new builds
- **On-premises vs. co-lo vs public cloud**

What's next in Sustainability?

Is Sustainability sustainable... ..yes, but...

- **Advancing AI methodologies to better optimize the energy, performance, carbon footprint algorithm**
 - Being applied in some fashion today
 - Anticipated to increase in use as more sophisticated models are developed
- **Necessity is the mother of invention**
 - Current projections of energy demand outstripping power production are increasingly alarming
 - Research currently being developed likely to advance into production to address both supply and demand
 - ARPA-E COOLERCHIPS*
 - Components relative to secondary cooling loop
 - Cooling systems for modular and edge data centers
 - Data center cooling system software
 - Testing facilities for items #1 & #2

*Cooling Operations Optimized for Leaps in Energy, Reliability and Carbon Hyper-efficiency for Information Processing Systems.

Questions?



**We welcome questions,
comments, and suggestions**

**Please contact us at:
mnosokoff@hyperionres.com**