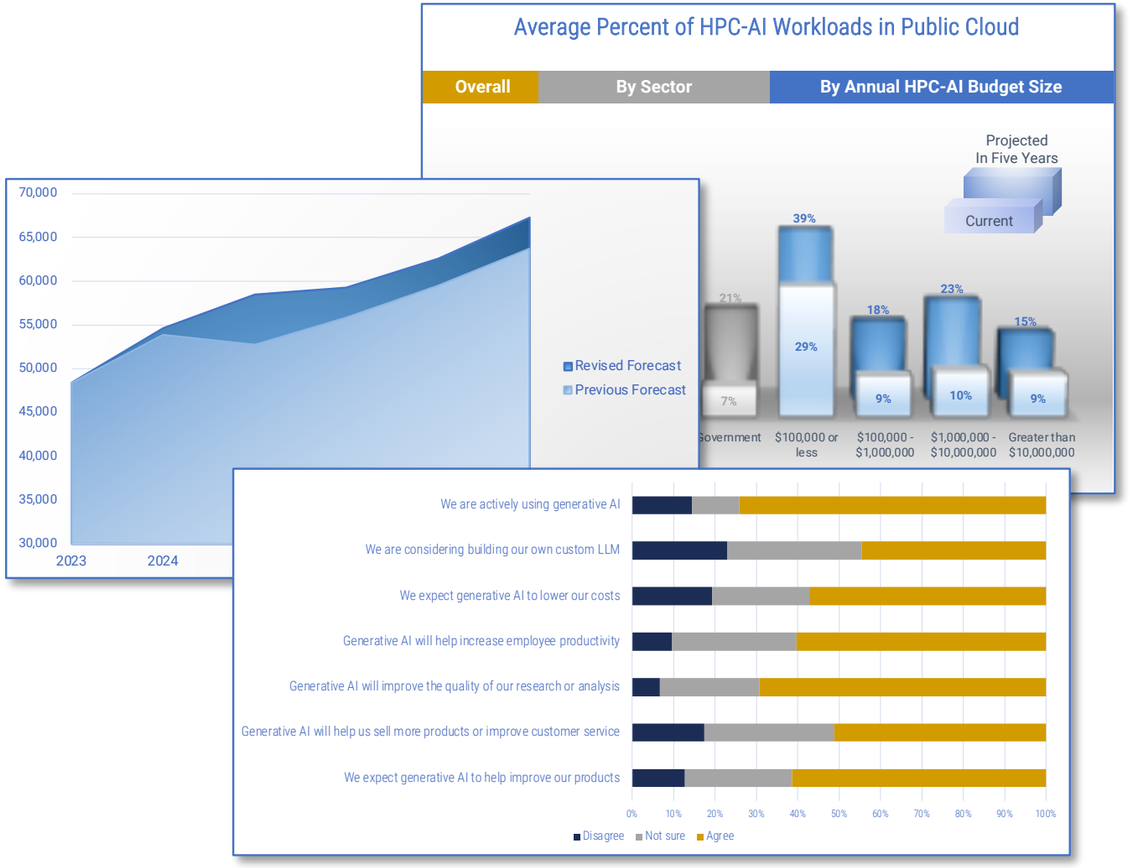
Five Big Questions for HPC-AI in 2025

Addison Snell, Intersect360 Research
addison@intersect360.com

Intersect360 Research 2025

- Now in 19th year tracking high-performance data center trends: HPC, AI, big data, cloud, hyperscale computing, etc.
- Market forecasts and trend analysis driven by end-user research
- Anchored by HPC-AI Leadership Organization (HALO), www.hpcaileadership.org



Intersect360 Research Team

Addison Snell
CEO, Co-Founder

Steve Conway
Advisor

Kevin Jackson
Analyst

Antonia Maar
Analyst

Chris Willard
Co-Founder, Advisor

Frank Richardson
Dir. Client Relations

Christine Fronczak
HALO Community Manager

Paul Muzio
Global HALO Facilitator

HPC-AI Leadership Organization (HALO)

- Global end-user organization for HPC and AI
- Help steer the industry by informing our research calendar and topics
- Free access to webinars, research, and members-only events
- No cost to participate – apply to join

HALO
HPC-AI LEADERSHIP ORGANIZATION

www.hpcaileadership.org

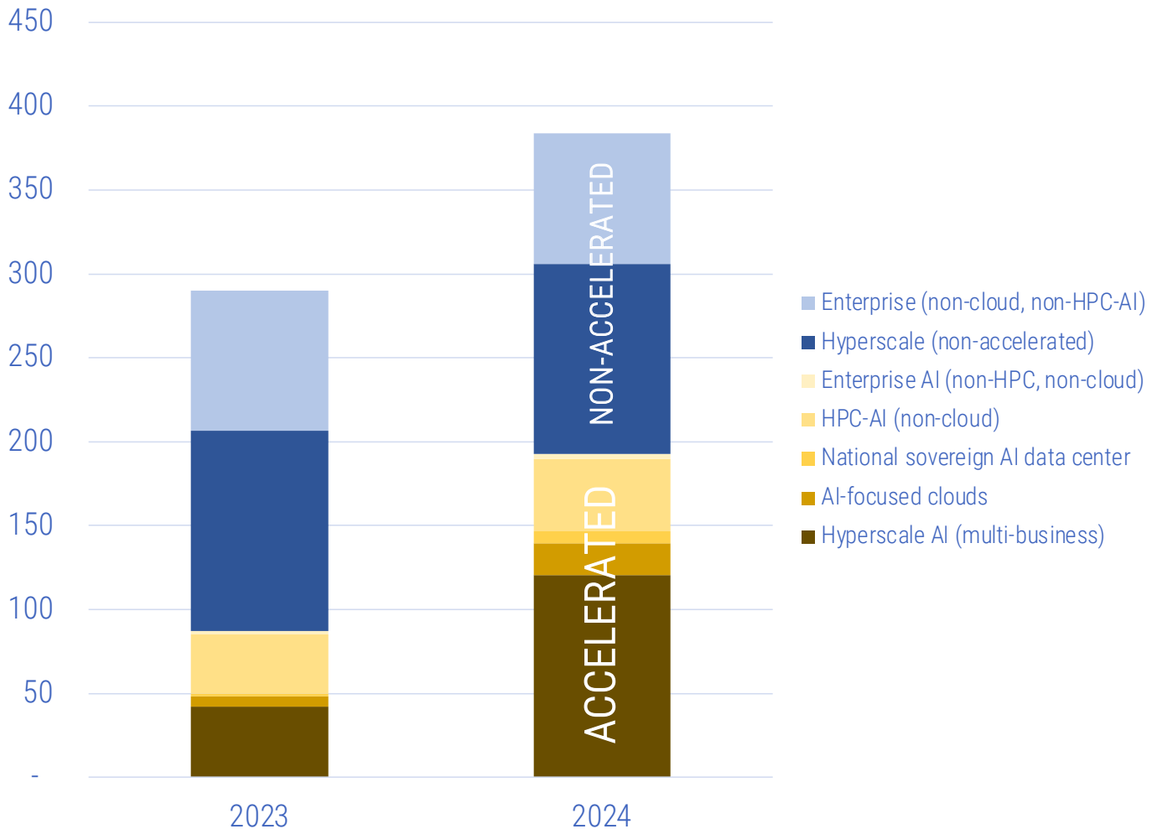
Question 1

How big can the AI market get?

Data Center Infrastructure Segmentation

- “AI market” could apply to anything in data center, and beyond. Here we focus on what data center segments are “accelerated” or “performance optimized.”
- Hyperscale infrastructure refers to companies with internet-driven business model in segments beyond cloud computing, such as digital content streamers, online retailers, social media sites, and online game hosting. Examples include Alibaba, Amazon, ByteDance, Meta, Microsoft, Oracle, and X. Hyperscale includes both accelerated AI infrastructure and non-optimized data center infrastructure.
- AI cloud service providers typically focus on serving AI workloads and do not have hyperscale business models other than cloud. Examples include CoreWeave, Lambda, and Nebius.
- National sovereign AI data centers are government-owned or government-controlled and focus on AI capabilities independent of HPC. Examples incloud G42 (UAE) and ABCI (Japan).
- HPC-AI (non-cloud) refers to infrastructure that might serve HPC or AI workloads. Many national lab “AI supercomputers” fit in this segment because they serve both HPC and AI. (There is not enough HPC without AI to warrant tracking.)
- Enterprise AI refers to organizational infrastructure exclusively focused on AI, not traditional HPC.

Data Center Infrastructure ($B), 2023 vs. 2024



Total data center infrastructure grew 32% year-over-year, driven entirely by accelerating computing segments (gold/yellow tones)

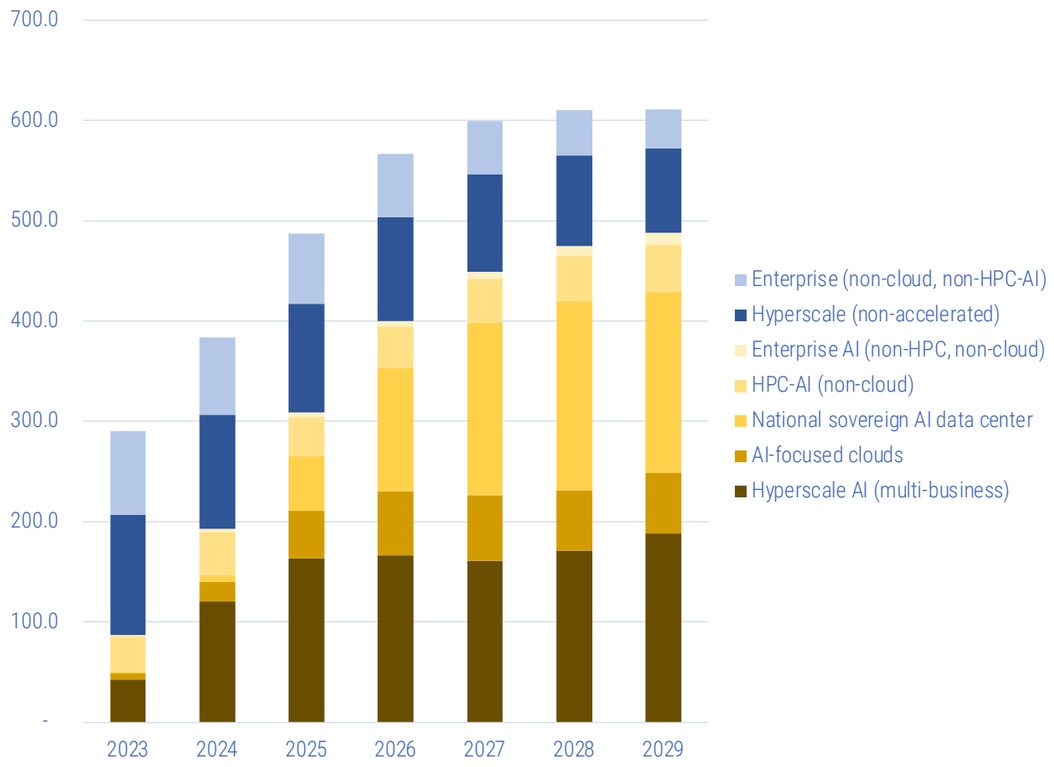
Accelerated computing is now 50% of all data center, up 121% year-over-year

Hyperscale AI grew 185% year-over-year, to over $120 billion

Hyperscale (accelerated and non-accelerated) accounts for over 60% of total data center spending

Despite growth, HPC is declining in share as a segment of the AI market

Data Center Infrastructure Forecast ($B)



Total Data Center Market, 9.7% CAGR

- Rapid growth through 2026, flattening in 2027
- Non-accelerated segments in decline

Accelerated segments, 20.4% CAGR

- Hyperscale AI levels off after 2025
- Ongoing growth in AI clouds and national sovereign AI initiatives (assuming U.S. will invest significantly)
- HPC-AI and Enterprise AI are a declining share of the accelerated computing paradigm

Enterprise HPC-AI Budget Research

Surveyed over 400 enterprises in U.S. and Europe with at least $10M annual revenue

92% had HPC, AI, or both as part of IT budget

Combined with annual HPC-AI Budget Map survey to inform new market forecast

HPC-AI Budgets

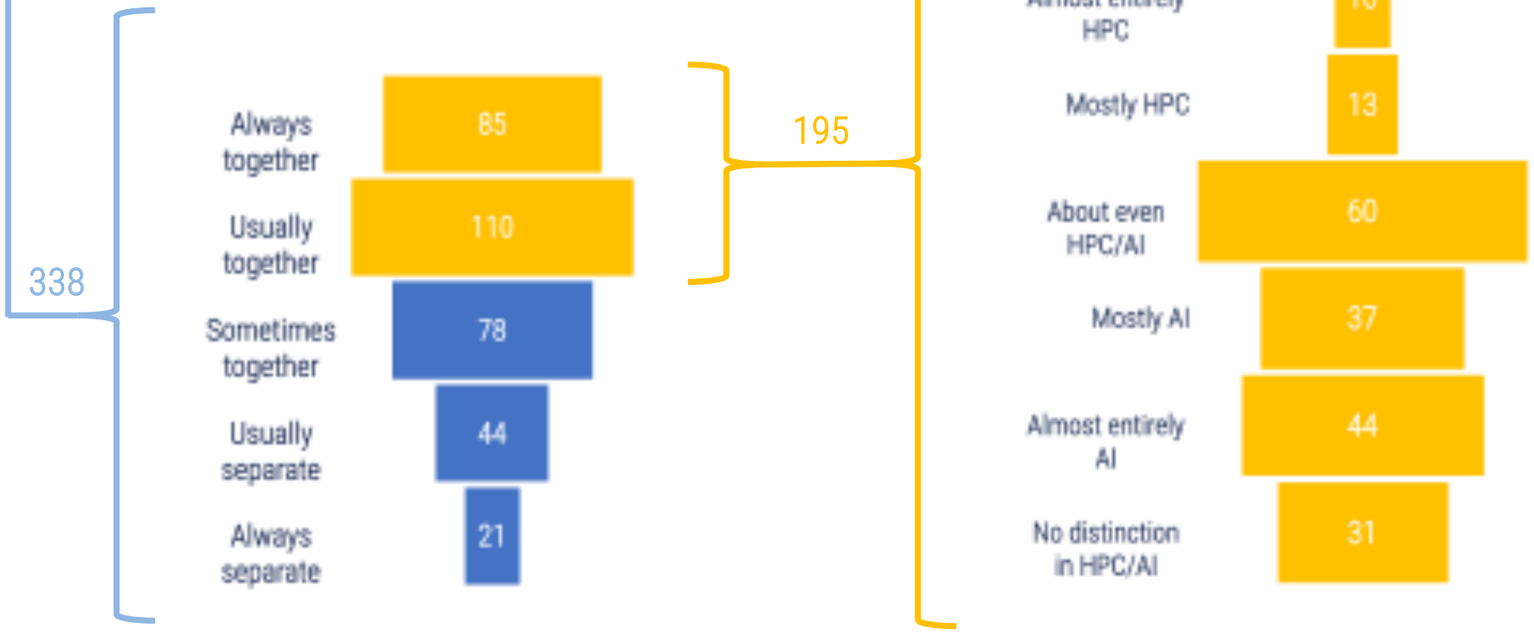
Among commercial (non-hyperscale) use cases, AI is usually found together with HPC

Almost all HPC users are now also AI

More often than not, HPC and AI are merged as part of the same budget

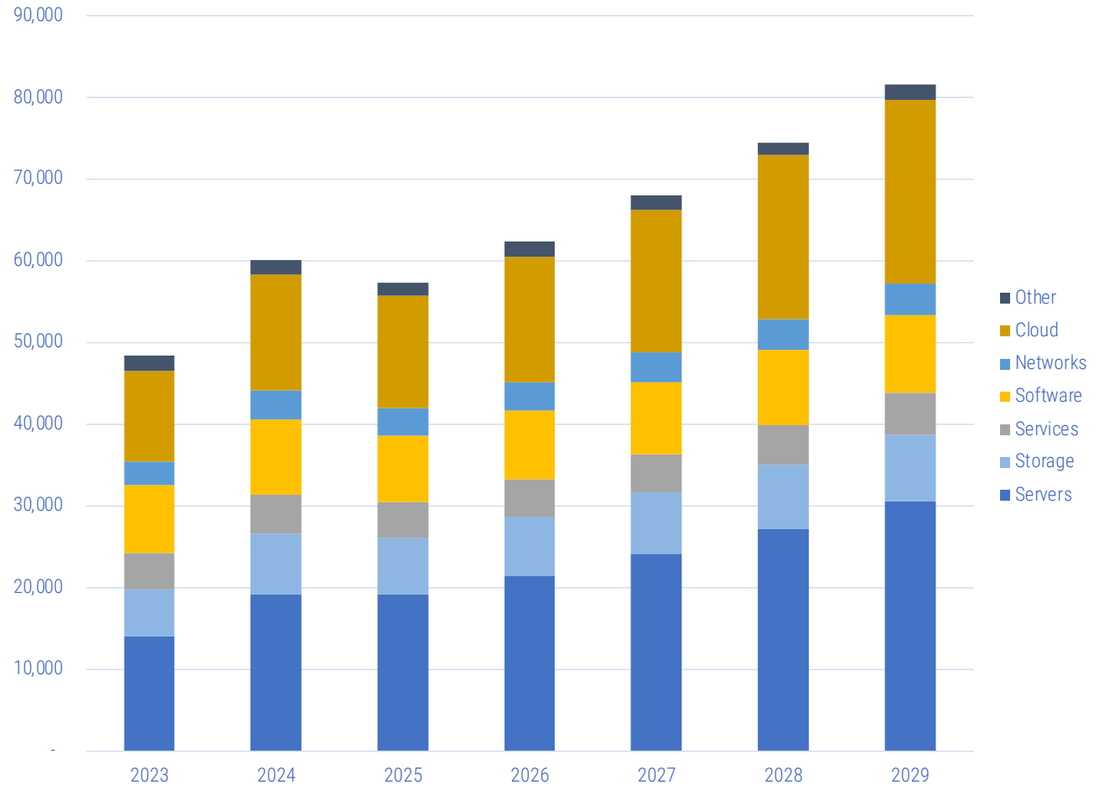
When merged, the budgets are more focused on AI than HPC

Budgets for HPC or AI	Count	Pct.
HPC only	15	3.8%
AI only	40	10.2%
Both HPC and AI	338	86.0%
Total	**393**	**100.0%**



HPC-AI and Enterprise AI Forecast ($M): Products and Services

- Market down -4.5% in 2025, due to funding cuts for U.S. academic grants and some other U.S. national agencies
- HPC-AI servers flat in 2025, capturing higher proportion of spending
- Market recovers to 6.3% CAGR despite down 2025
- Highest CAGRs in Servers, Cloud
- Cloud could resume ultra-high growth in 2030s if hyperscale model prevails over on-prem



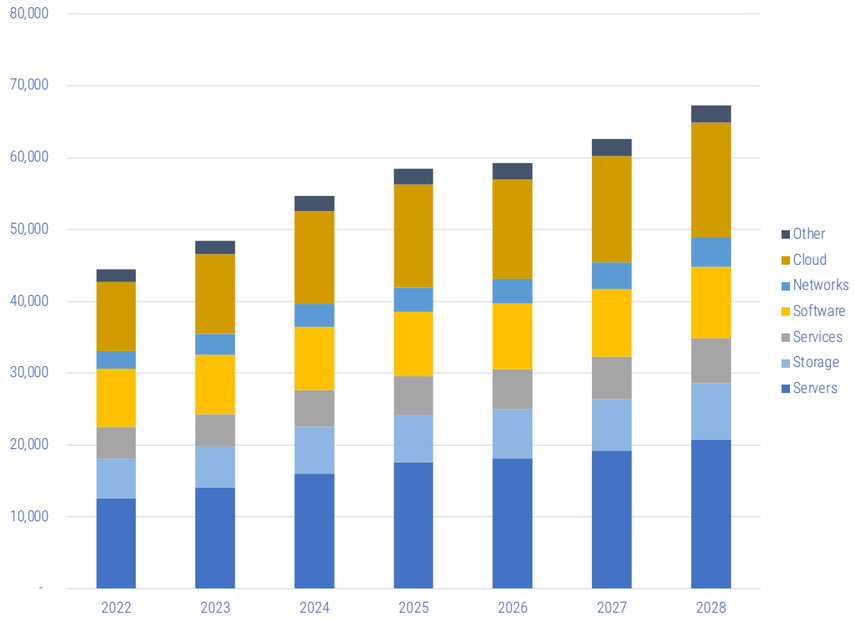
“Enterprise AI” Opportunity

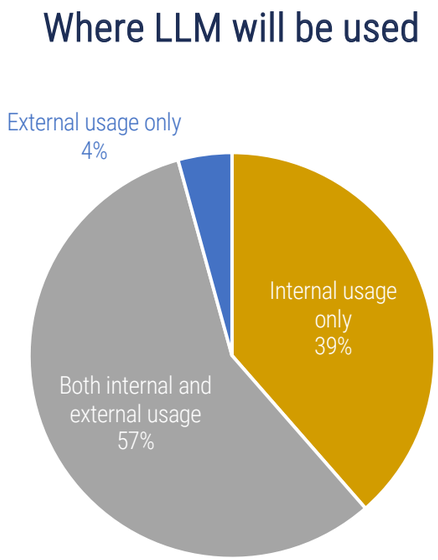
- Two paths to profitable investment: 1. Increase revenue. 2. Decrease cost.
- Most of the focus has been on costs: operational efficiency, reduced headcount, etc.
  - How much money will you spend to save $100?
  - Diminishing returns at scale
- Two paths to increasing revenue: 1. Larger overall market. 2. Steal share from competitor.
  - What markets actually get bigger because of AI?
  - Stealing share is zero-sum game that leads to prisoner’s dilemma scenarios. AI is new “cost of doing business.”
- AI for science or R&D is different from consumer or horizontal enterprise AI

Question 2

Will hyperscale completely take over enterprise computing?

Cloud Penetration in HPC-AI



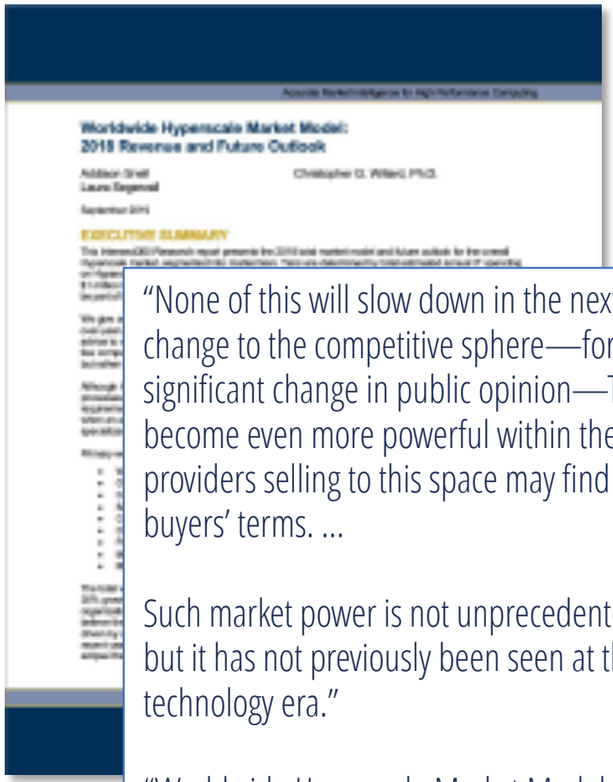


Cloud has been approaching an asymptote of penetration in HPC-AI …

But what if cloud is the only choice?

The Power of Hyperscale

- Nearing three-quarters of data center spending worldwide
- Hundreds of megawatts, even gigawatts, at a time
- Technology vendors prioritizing deliveries


Worldwide Hyperscale Market Model:
2018 Revenue and Future Outlook

Addison Snell
Laura Segenal

Christopher G. Willard, Ph.D.

September 2019

EXECUTIVE SUMMARY

"None of this will slow down in the next few years. Barring a significant change to the competitive sphere—for example, through regulation or significant change in public opinion—Tier 1 Hyperscale companies will become even more powerful within the forecastable timeframe. Technology providers selling to this space may find they have to do it strictly on the buyers' terms. …

Such market power is not unprecedented in the world's economic history, but it has not previously been seen at this level in the information technology era."

"Worldwide Hyperscale Market Model: 2018 Revenue and Future Outlook," September 2019

Question 3

What effect will the new U.S. administration have on HPC-AI?

(including European AI Factories)

HPC-AI Nationalism and the Role of Government

HALO
HPC-AI LEADERSHIP ORGANIZATION

HiPEAC

HALO and HiPEAC both highlighted HPC-AI nationalism issues as threats to progress

VS.

National Initiatives

From the Introduction:

"Winning the AI race will usher in a new golden age of human flourishing, economic competitiveness, and national security for the American people. AI will enable Americans to discover new materials, synthesize new chemicals, manufacture new drugs, and develop new methods to harness energy—an industrial revolution."

"We need to build and maintain vast AI infrastructure and the energy to power it. To do that, we will continue to reject radical climate dogma and bureaucratic red tape, as the Administration has done since Inauguration Day. Simply put, we need to 'Build, Baby, Build!'"

EuroHPC AI Factories

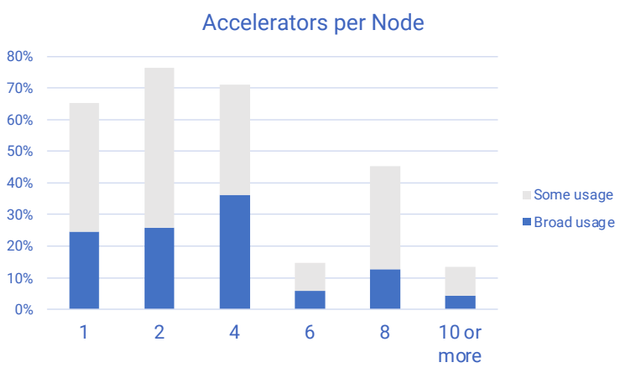
- Under EuroHPC JU, focus is on overlap with science and R&D
- “Expanding and scaling business innovation for SMEs and Startups”
- Will help to develop and scale European technologies for HPC and AI
- Promotes European values of data privacy and sustainable technologies

Question 4

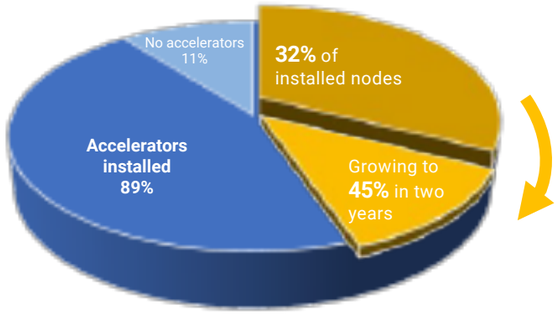
Can anyone challenge Nvidia?

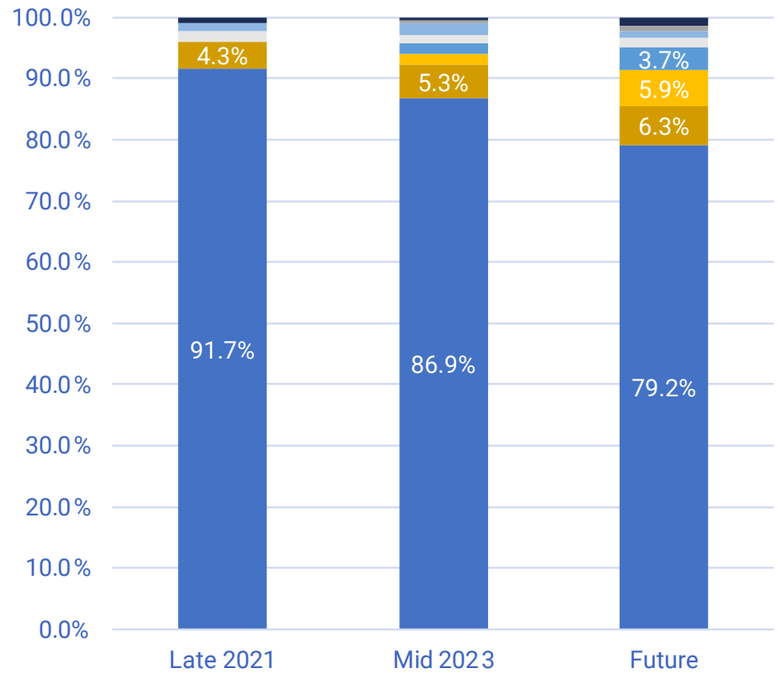
Usage of Accelerators

Accelerators in HPC-AI



Four GPUs / node remains the most common configuration, "balancing" technical computing and AI





AMD share (APU + discrete GPU) projected to double, but Nvidia still dominates

"Late 2021" represents previous survey iteration. "Future" is current survey respondents' expectation of usage in two to three years.

Potential Challengers to Nvidia

Conventional

intel

AMD

Startup

cerebras

groq

SambaNova SYSTEMS

InspireSemi™

GRAPHCORE

New Paradigm

Azure

aws

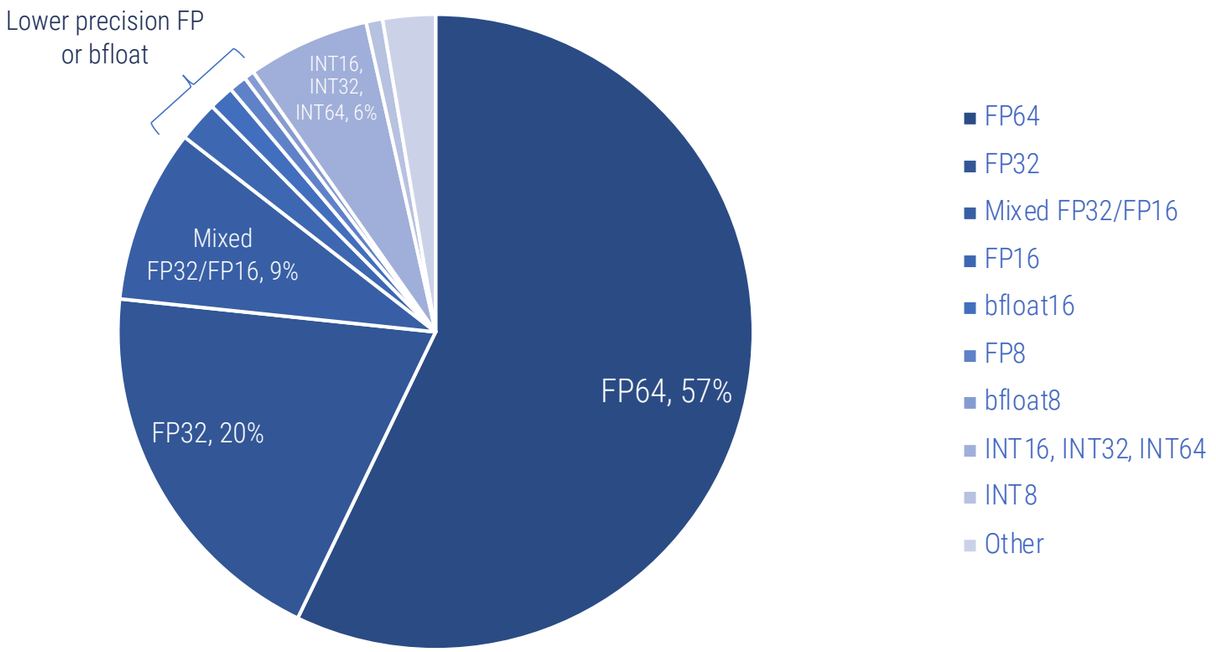
xAI

Google Cloud

Question 5

What about good old HPC?

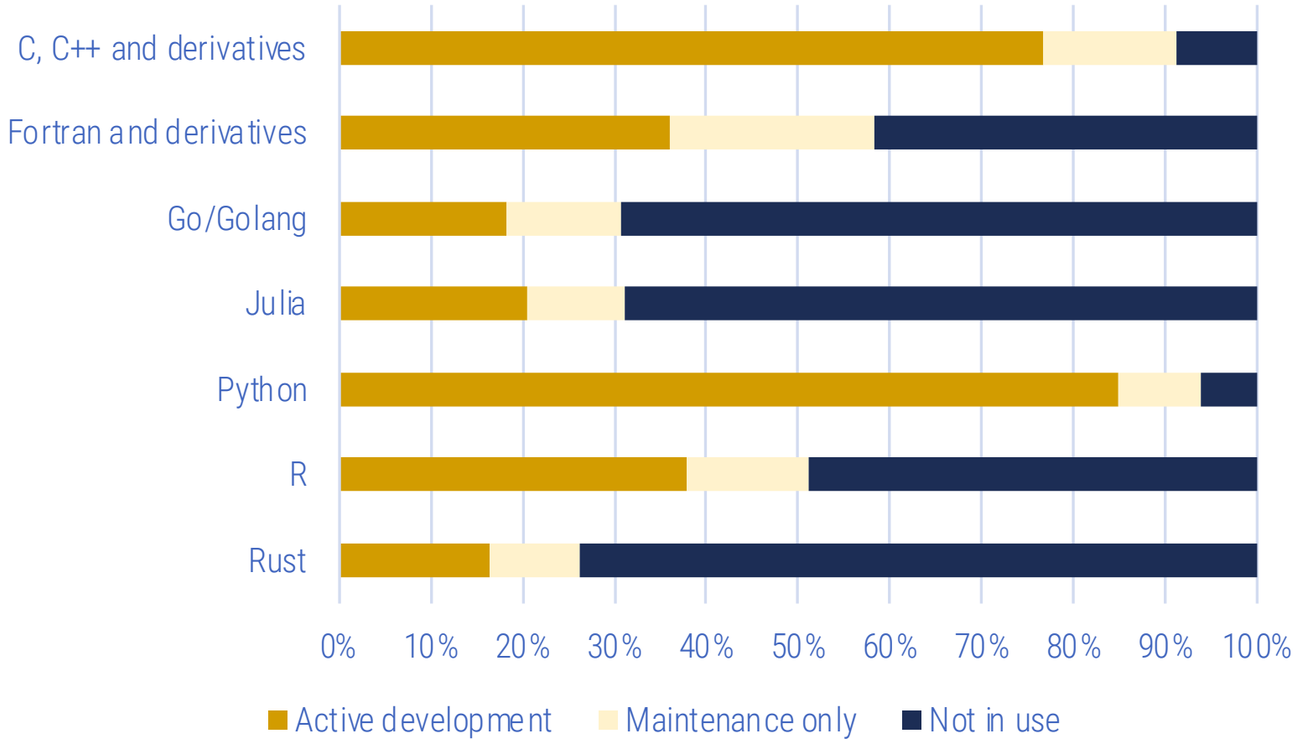
Levels of Precision



Weighted averages based on total respondents in each domain

- Not everything requires 64-bit
- Highest proportion of FP64:
  - Chemistry, 72%
  - Astrophysics, physics, weather, 65%
- Low-precision FP, bfloat are rare
- Highest proportion of INT (all):
  - Visualization, 13%
  - Biosciences, 12%
  - Finance, 11%

Programming Languages



Ignoring "not sure" responses

- Buoyed by AI revolution, Python has become a dominant language for HPC-AI
- C/C++ still very common
- Fortran still has an important role but is (very) slowly fading into maintenance

HPC and AI: Convergence or Divergence?

- We've talked about convergence of HPC and AI for years, but there are signs that they are pulling apart in some ways
- AI is driving funding
- Processors and systems (compute side) are focusing on AI in ways that might not serve HPC
- Storage and networking seems to be more compatible in terms of investment (so far)

Have We Heard This Song Before?

That’s not real HPC

Can’t solve the hard problems

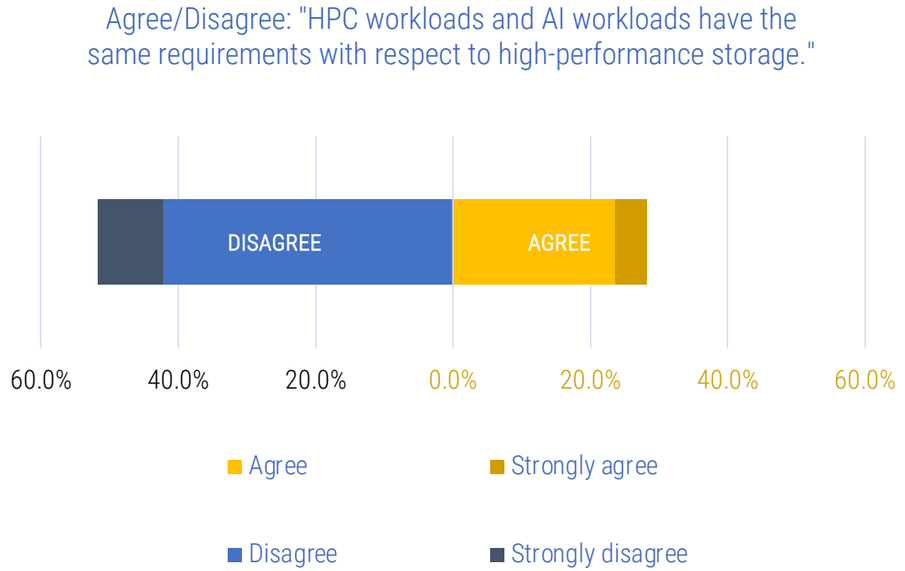
Flops don’t translate to real application performance

“State of the HPC-AI Market” Reports

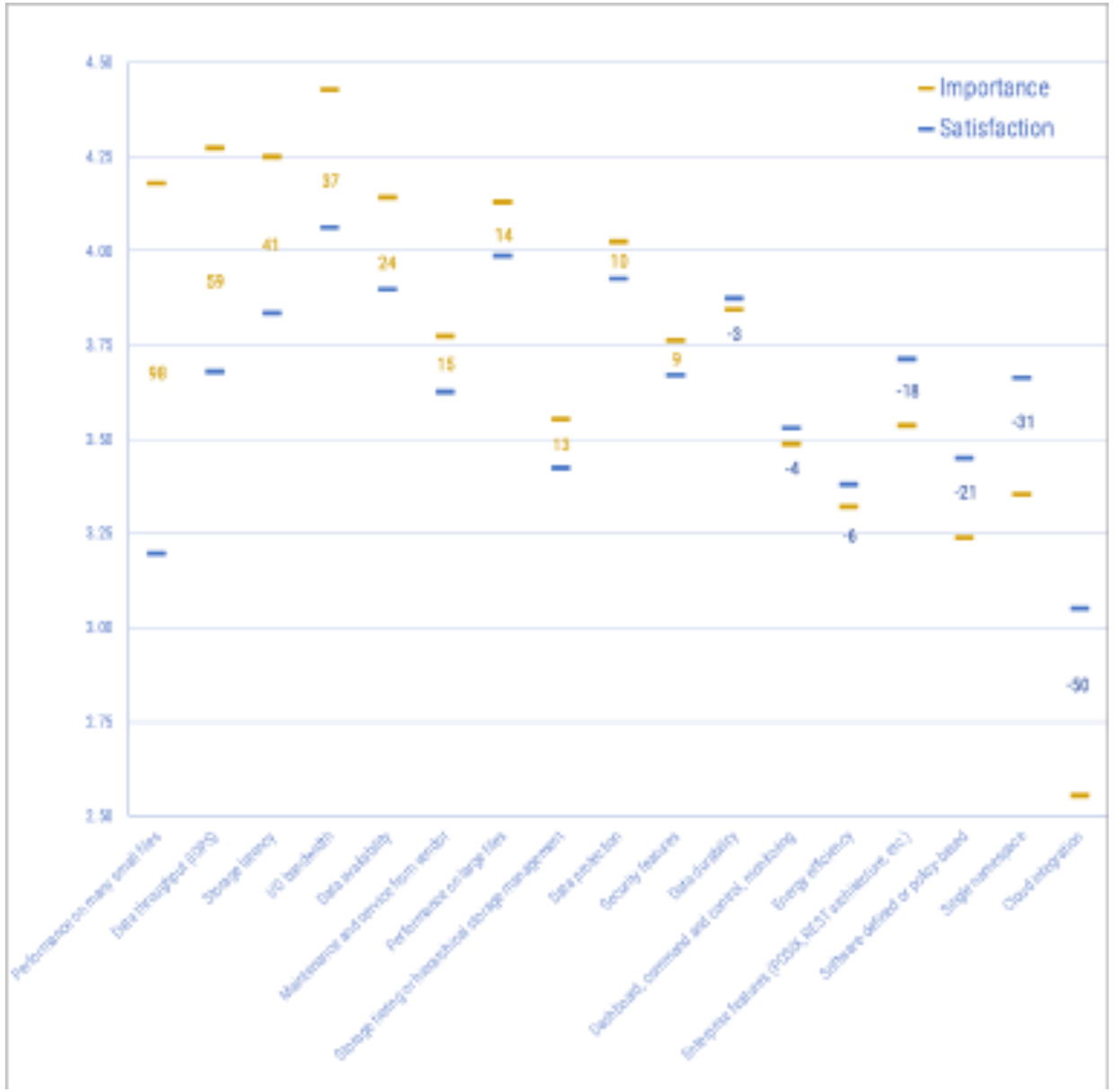
- Divided into technology modules
  - Processing elements (CPUs, GPUs, etc.)
  - Quantum computing
  - Systems
  - Interconnects and networking
  - Storage and data management
  - Cooling and facilities
  - Cloud computing
  - Other topics by demand
- HALO end user surveys:
  - Planned adoption of new technologies
  - Importance of technology features
  - Satisfaction with current solutions
  - Gap analysis,
- Inputs from key suppliers:
  - Target applications served
  - Key differentiation
  - Future outlook

State of the HPC-AI Market: Storage / Data Mgmt

	HPC only	AI only	Mixed HPC-AI	Total
More than 1 EB	2	1	3	6 (4%)
100 PB – 1 EB	6	2	11	19 (12%)
10 PB – 100 PB	8	6	25	37 (24%)
1 PB – 10 PB	18	15	30	63 (41%)
Less than 1 PB	6	12	12	30 (19%)



93% of survey respondents say "storage solutions are strategically important, carefully considered acquisitions."



Satisfaction Gaps: Storage / Data Mgmt

Survey in progress for networking and interconnects ...

Advice on AI and Quantum Adoption

- For AI, know what your goal is – lower cost, greater productivity, product innovation?
  - Do you have the ingredients (data, skills, infrastructure) to support the goals?
  - When are specialized tools needed, versus general ones?
  - Is it strictly internal, or visible to customers and partners?
  - Carefully consider on-prem vs. cloud in terms of cost, capability, and data locality
  - Do you need to lead, or can you follow?
- Quantum technologies are available but not mainstream. Consider cloud access, quantum emulation, and quantum acceleration technologies
- Have partners who understand your specific problems, more than general enterprise

Five Big Questions for HPC-AI in 2025

Addison Snell, Intersect360 Research
addison@intersect360.com