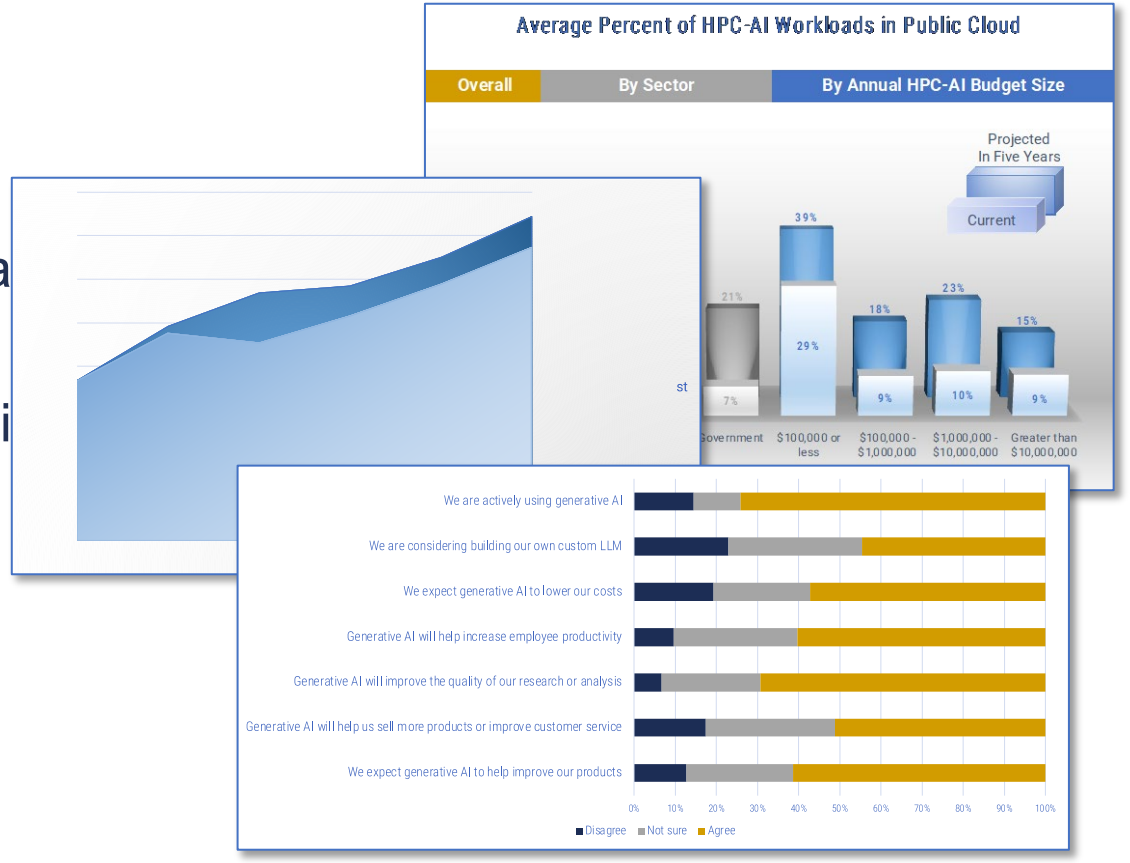
Five Big Questions for HPC-AI in 2025

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 AI Leadership Organization (HALO), www.hpcaileadership.org



Intersect360 Research Team

Addison Snell
CEO, Owner

Steve Conway
Senior Analyst

Kevin Jackson
Analyst
New hire!

Antonia Maar
Analyst
New hire!

Frank Richardson
Dir. Client Relations

Kara Ketchum
Marketing Associate
New hire!

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 events for members
- No cost to participate – apply to join

HALO
HPC-AI LEADERSHIP ORGANIZATION

www.hpcaileadership.org

Question 1

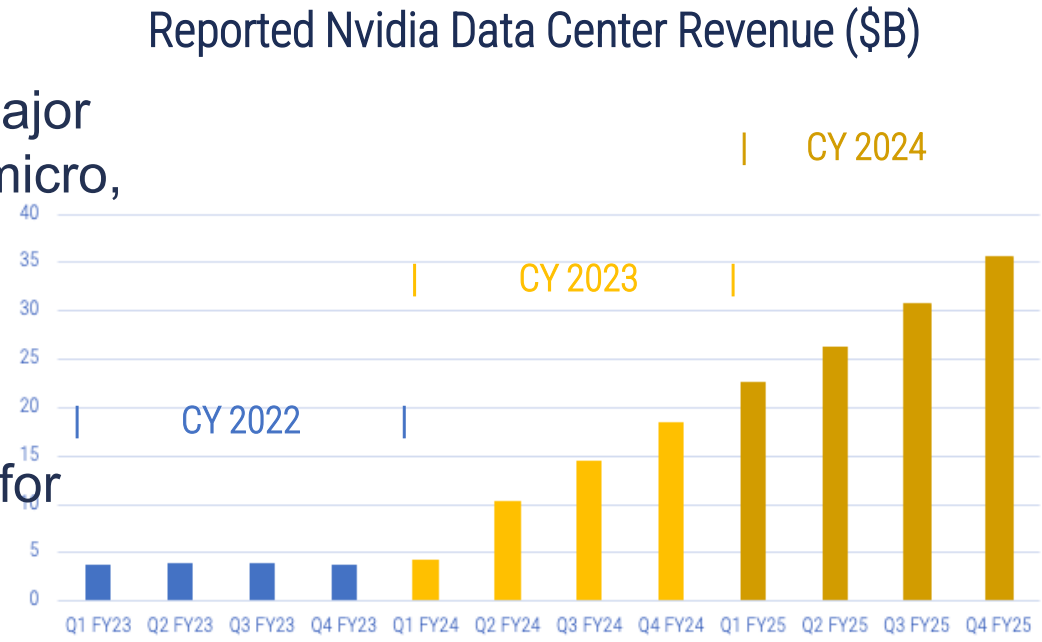
How big can the AI market get?

HPC-AI Market: Mid-2024 update

- All major suppliers are trending well above forecast for 2024.
- Continued exponential growth in hyperscale AI is the primary driver, exce
  - xAI became an unexpected top-tier competitor with Amazon, Google, Meta, Mic
  - Top hyperscale companies now spending in excess of $10B per year on AI infrastructure
  - Base metric for data centers is how many hundreds of megawatts they
- Additionally, on-premises AI is beginning to take off. This would look like a ma it not dwarfed by hyperscale spending.
- Forecasted pause in market growth slides from 2025 into 2026.

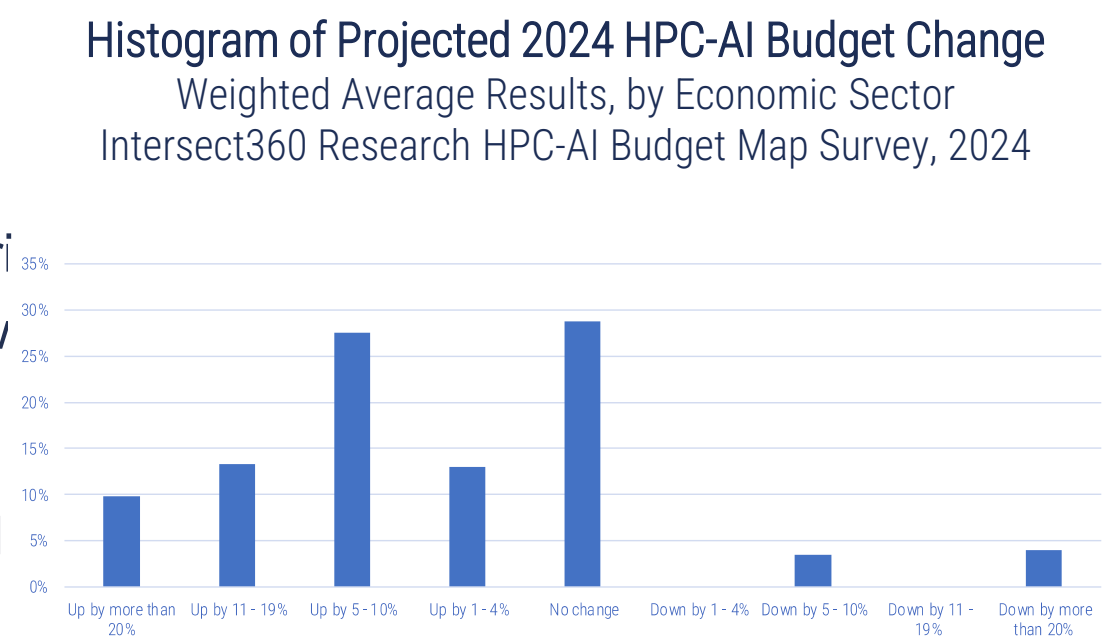
HPC-AI Supplier Analysis

- Midyear check on 2024 revenue for major suppliers, including HPE, Dell, Supermicro, Lenovo, Nvidia, Intel, AMD, ...
- Most were trending to 75% to 150% growth
- HPE and Dell are usually bellwethers for on-premises HPC, both recognized major hyperscale AI revenue in 2024



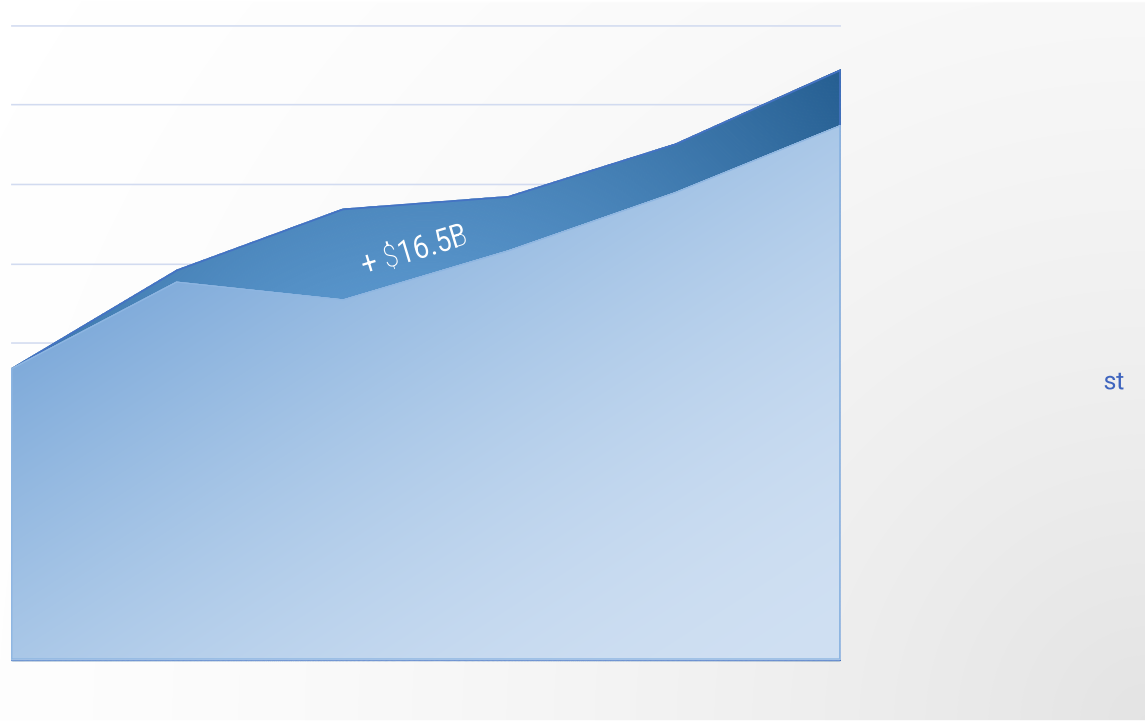
HPC-AI Budget Expectation Data Roll-Up

- Traditional HPC user database
  - Commercial, +8.3%
  - Blended market, +6.2%
- Separate survey of large enterpri
  - Overall, +8.6% (consistent w Intersect360 Research HPC survey database)
  - Larger budgets trend toward higher growth
  - Pure AI budgets slightly more growth than HPC-directed budgets



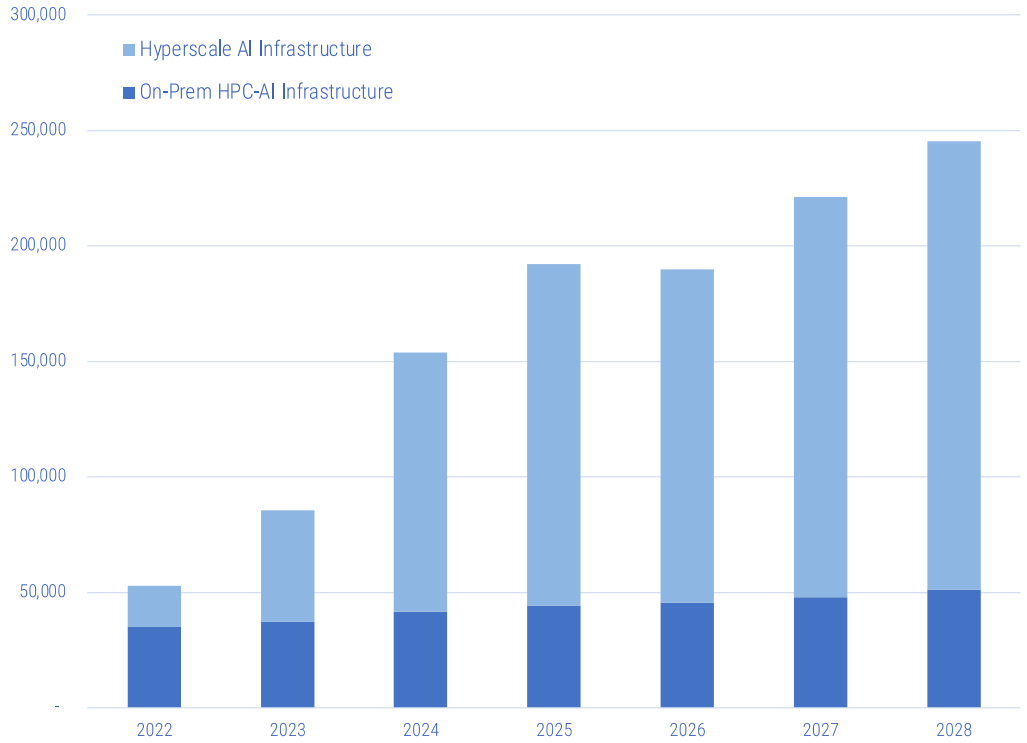
Revised On-Prem HPC-AI Forecast ($M)

- Slight increase to outlook fo this year
- Biggest difference is in 202 outlook, primarily due to on premises enterprise AI
- $16.5B in added revenue over five year span
- Five-year CAGR upgraded to 6.8% (was 5.6% in May 2024 forecast)



Revised HPC-AI Infrastructure Forecast ($M)

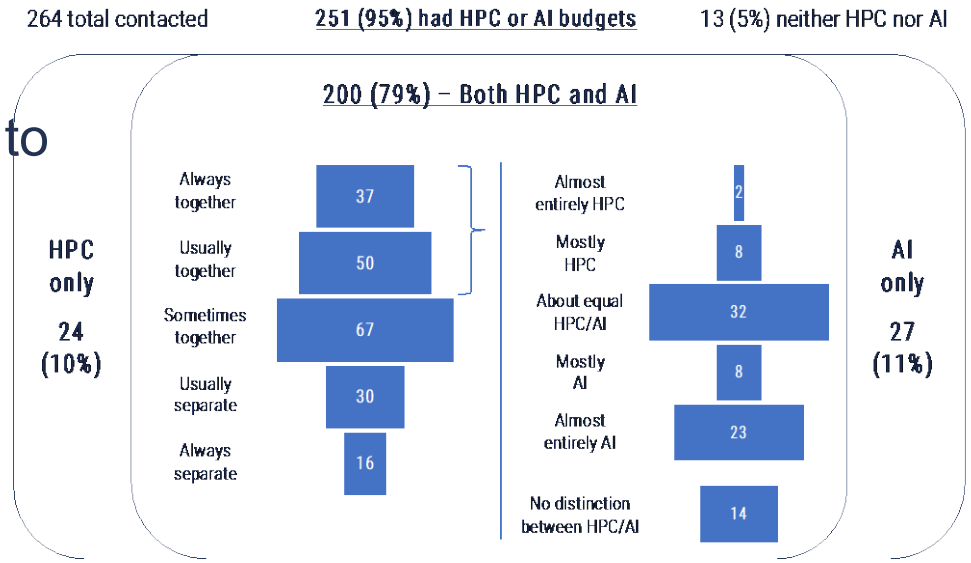
- Hyperscale AI has seen second straight year of triple-digit growth
- Hyperscale AI in 2024 is more than 6x where it was in 2022
- Hyperscale AI segment will near $200 billion in 2028
- On-premises HPC-AI infrastructure now forecast to grow 11.8% in 2024 (was 11.0% in May 2024 forecast)
- Increase in on-premises enterprise AI is dwarfed by growth in hyperscale



HPC-AI Budget Survey

- In process now, to be completed by end of March 2025
- Includes our traditional HPC survey list along with general enterprise computing to find penetration rates of HPC and AI
- Relationships between HPC and AI
  - Together or separate?
  - Relative growth rates?
- Data helps form AI HPC market forecast

From 2024 HPC-AI Budget Map Survey



Storage/Networking: HPC-AI vs. Hyperscale AI

On-Prem HPC

- Storage is about 25% of expenditure
- Tiered hybrid storage: flash, disk, tape
- Irregular use of parallel file systems – Lustre, GPFS, BeeGFS, custom, or often none
- HPC storage specialists share market with enterprise storage mainstays
- About half InfiniBand, half Ethernet – InfiniBand for larger, faster systems

Hyperscale AI

- Storage is about 5% of expenditure
- All flash
- "Data platforms" embed parallel file system usage
- HPC storage specialists have dominated b focusing on AI
- Nvidia networking (InfiniBand) has inside lane; UEC is an interesting development

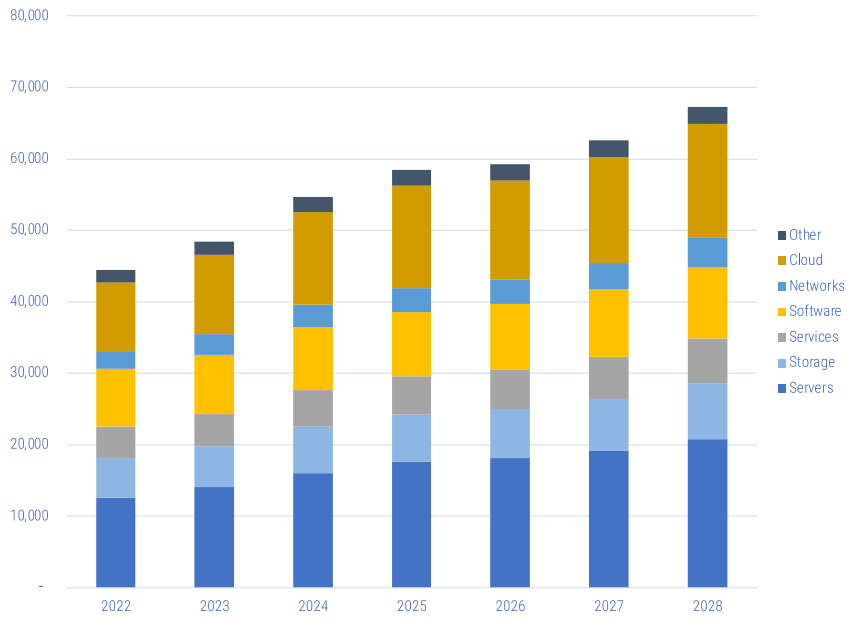
“Enterprise AI” Opportunity

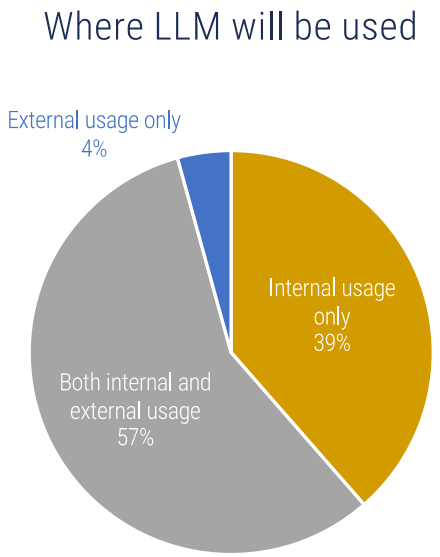
- Two paths to profitable investment: 1. Increase revenue. 2. Decrease cos
- Most of the focus has been on costs: operational efficiency, reduced head
  - How much money will you spend to save $100?
  - Diminishing returns at scale
- Two paths to increasing revenue: 1. Larger overall market. 2. Steal share
  - What markets actually get bigger because of AI?
  - Stealing share is a zero-sum game that leads to prisoner’s dilemma scenarios. “cost of doing business.”

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.

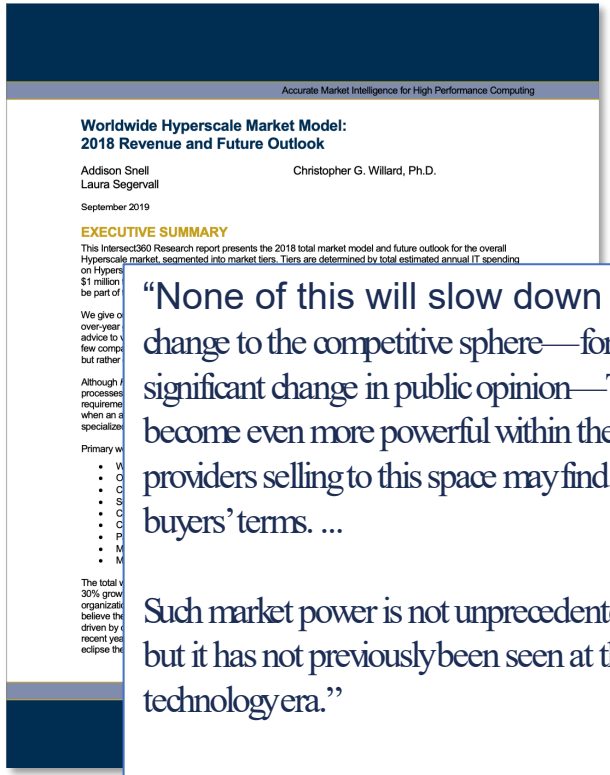
But what if cloud is the only choice?

The Power of Hyperscale

Over three-quarters of data center spending worldwide

Hundreds of megawatts, even gigawatts, at a time

Technology vendors prioritizing deliveries


Accurate Market Intelligence for High Performance Computing

Worldwide Hyperscale Market Model: 2018 Revenue and Future Outlook

Addison Snell
Laura Segervall

Christopher G. Willard, Ph.D.

September 2019

EXECUTIVE SUMMARY

This Intersect360 Research report presents the 2018 total market model and future outlook for the overall Hyperscale market, segmented into market tiers. Tiers are determined by total estimated annual IT spending

"None of this will slow down in the next few years. Barring a signif 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?

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.

NIH

NASA

NOAA

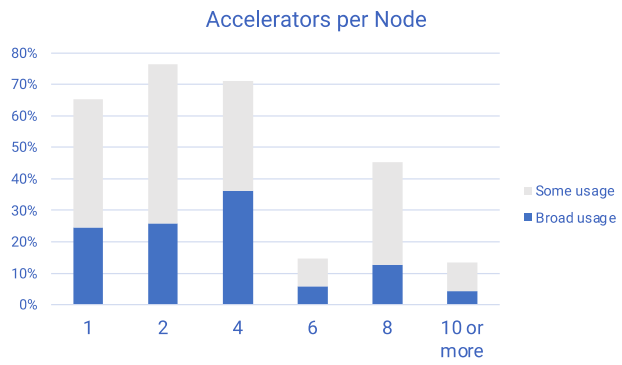
NSF

Question 4

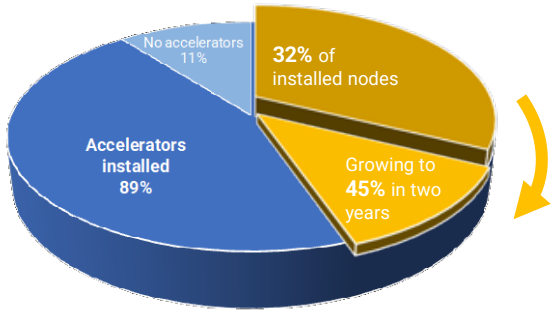
Can anyone challenge Nvidia?

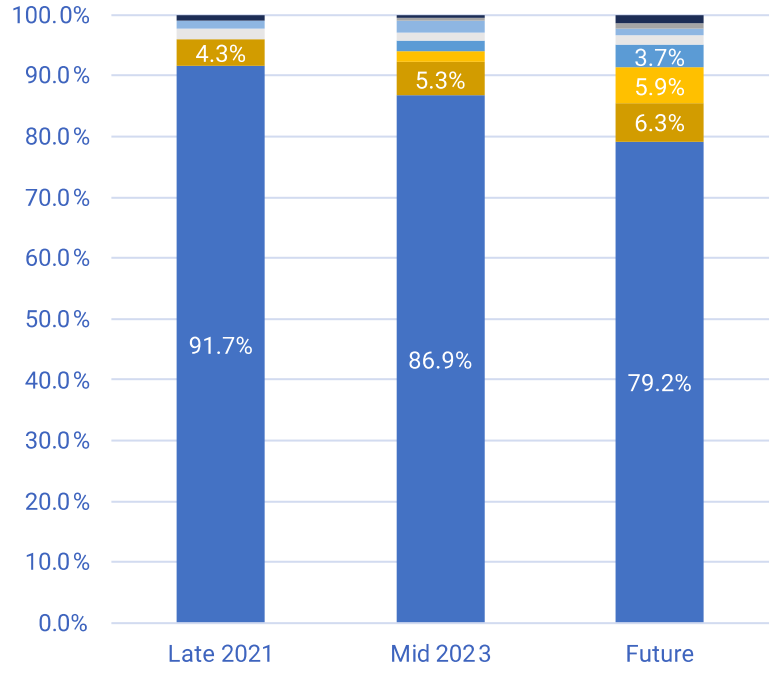
Usage of Accelerators

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

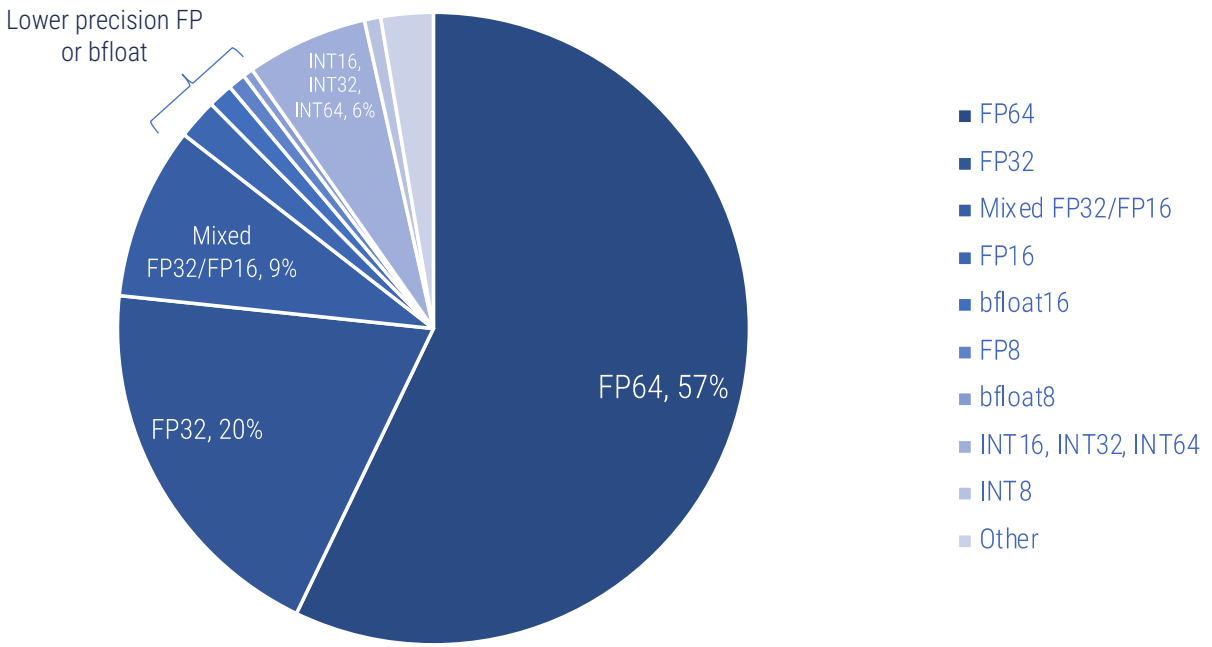
New Paradigm

- Azure
- aws
- xAI
- Google Cloud

Question 5

What about good old HPC?

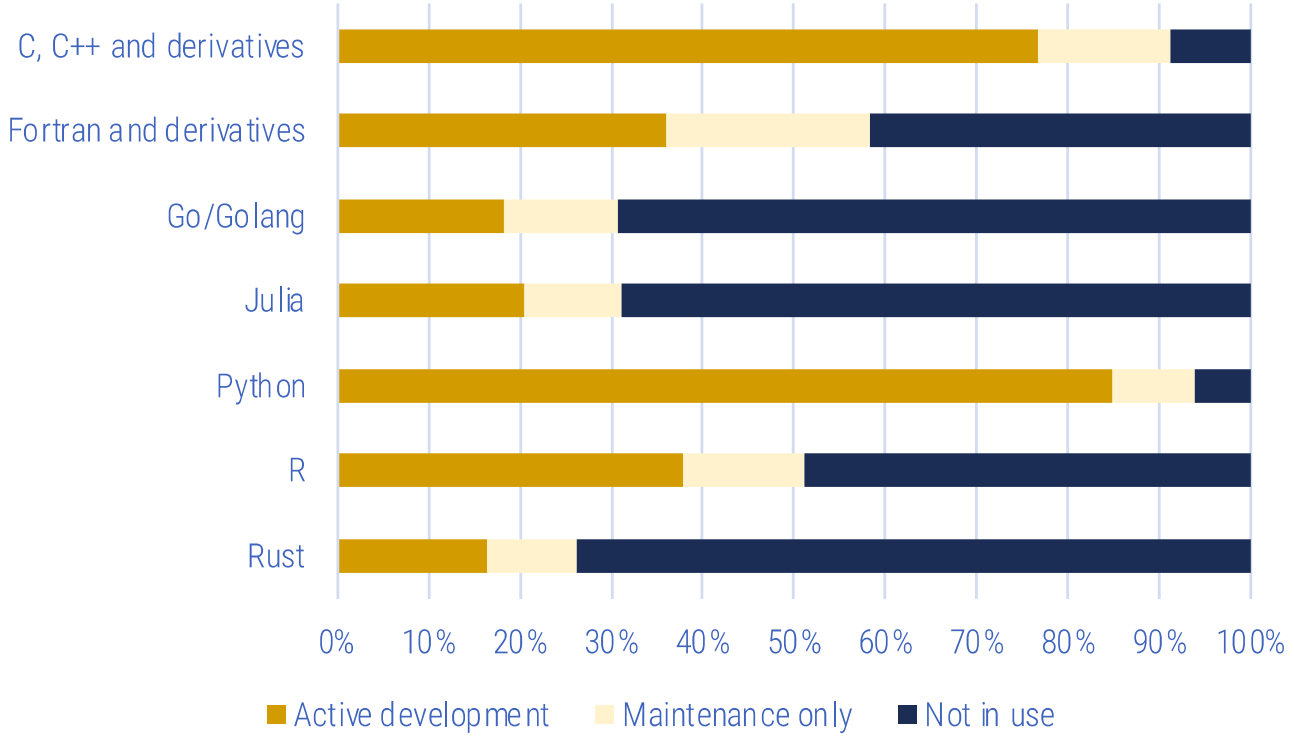
Levels of Precision



Weighted averages based on total respondents in each domain

- Not everything requires 64 bits
- Highest proportion of FP64:
  - Chemistry, 72%
  - Astrophysics, physics, weather, 65%
- Lower 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 s pulling apart in some ways
- AI is driving funding
- Processors and systems (compute side) are focusing on AI in ways that n
- Storage and networking seems to be more compatible in terms of investm

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: Planned Publication Schedule

		Deadline for Supplier Content	Planned Publication
1	Storage and Data Management	May 9	Early June
2	Systems	May 30	Early July
3	Cooling/Facilities	June 20	Late July
4	Quantum Computing	July 1	Early August
5	Interconnects and Networking	July 18	Late August
6	Processing architectures (CPUs, GPUs, etc.)	August 1	Early September
7	Cloud Computing	August 15	Late September

HPC-AI Research Calendar: Summary

- January – March: HPC budget surveys. Insights provided to clients; not publis
- April – May: HPC market forecasts
- June – September: State of the HPC Market reports
  - June: Storage and data management
  - July: Systems; cooling and facilities
  - August: Quantum computing; interconnects and networking
  - September: Processing architectures; cloud computing
- Ongoing and by end of year: Additional reports on identified topics of inter
  sustainability, AI, national sovereignty, ethics in AI

Research fueled by:

HALO
HPC-AI LEADERSHIP ORGANIZATION

Five Big Questions for HPC-AI in 2025

Addison Snell, Intersect360 Research
addison@intersect360.com