Pre-SC24 HPC-AI Webinar

HPC-AI Market Forecast Update
Insights from HPC-AI Leadership Organization (HALO) Interviews:
Issues Facing the HPC-AI Industry

November 2024

Intersect360 RESEARCH

©2024 Intersect360 Research

Agenda

- Presentation (60 minutes)
  - New: Webinar contains more data, with exclusive access for paid clients and HPC-AI end-user members of HPC-AI Leadership Organization (HALO)
  - Update to HPC-AI 2024-28 market forecast
  - Insights from first HALO report, based on interviews with global HALO Advisory Committee members: Issues Facing the HPC-AI Industry
  - Where to find us and what to watch for at SC24
- Live Q&A (30 minutes)

©2024 Intersect360 Research

Worldwide HPC and AI Market

2024 midyear update, revised 2024-28 forecast

October 2024

Intersect360 RESEARCH

©2024 Intersect360 Research

What Is the HPC-AI Market?

Intersect360 Research has a proven history tracking high-performance, scalable data center technologies

Included: Multi-node, networked systems or cloud instances running parallel applications, generally requiring focus on performance or scalability in some dimension (e.g. processing, memory, I/O, networking); plus associated storage, software, services, etc.

Not included: Single-node desktops or workstations; embedded or edge devices

©2024 Intersect360 Research

On-Premises vs. Hyperscale Infrastructure

Hyperscale AI Infrastructure

- Companies are those with internet-driven business models, such as cloud service providers, digital content streamers, online retailers, social media sites, and online game hosting
- Spending hundreds of millions to billions of dollars per year in total IT infrastructure
- Amazon, Apple, Google, Meta, and Microsoft are largest examples, but also includes non-U.S. companies such as Alibaba, ByteDance, Naver, and Tencent, as well as AI-focused cloud services such as CoreWeave, Denvr Dataworks, and Lambda

On-Premises HPC-AI Infrastructure

- All HPC-AI infrastructure outside of the relatively few hyperscale companies
- Multi-node, networked systems or cloud instances running parallel applications, which require a focus on performance or scalability in some dimension
- Includes infrastructure exclusive to HPC or AI
- "On-premises" includes remote or co-located data center infrastructure; really means "not in the cloud"
- "Infrastructure" does not include cloud spending; ignored in this segmentation to avoid double count

©2024 Intersect360 Research

Midyear Forecast Methodology

- Analysis of major vendors' year-to-date revenue
- Includes analysis of on-premises HPC-AI versus hyperscale AI data centers
- Primary segmentation: products and services (servers, storage, software, services, networking, cloud, other)
- Other segmentations are back-calculated assuming the proportionality from previous forecast remains unchanged
- Refer to forecast published May 2024 for full methodology

©2024 Intersect360 Research

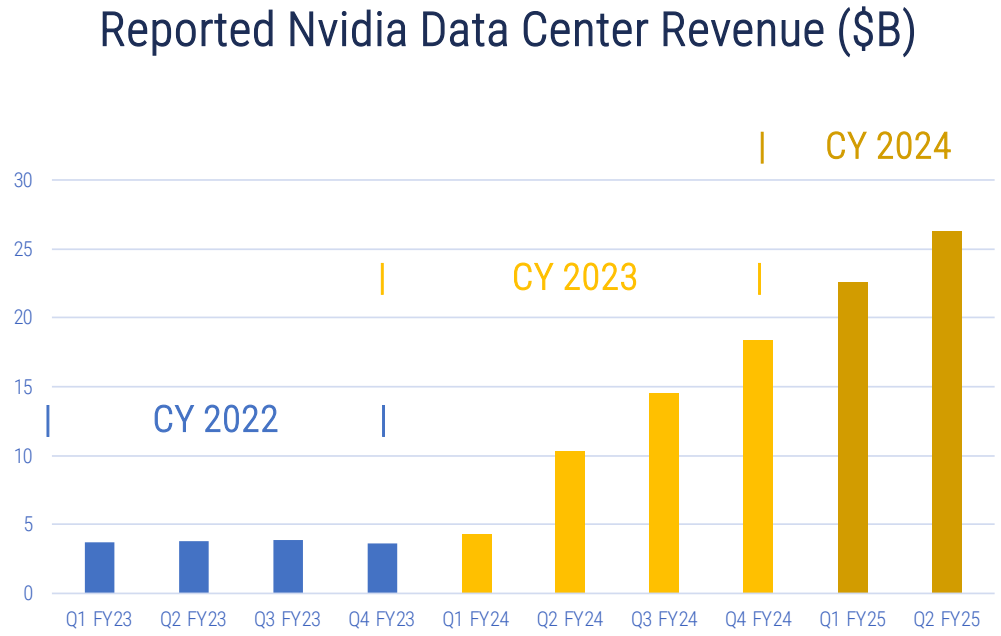
HPC-AI Market: Mid-2024 update

- All major suppliers are trending well above forecast for 2024.
- Continued exponential growth in hyperscale AI is the primary driver, exceeding forecast:
  - xAI became an unexpected top-tier competitor with Amazon, Google, Meta, Microsoft, …
  - 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 consume
- Additionally, on-premises AI is beginning to take off. This would look like a major trend were it not dwarfed by hyperscale spending.
- Forecasted pause in market growth slides from 2025 into 2026.

©2024 Intersect360 Research

HPC-AI Supplier Analysis

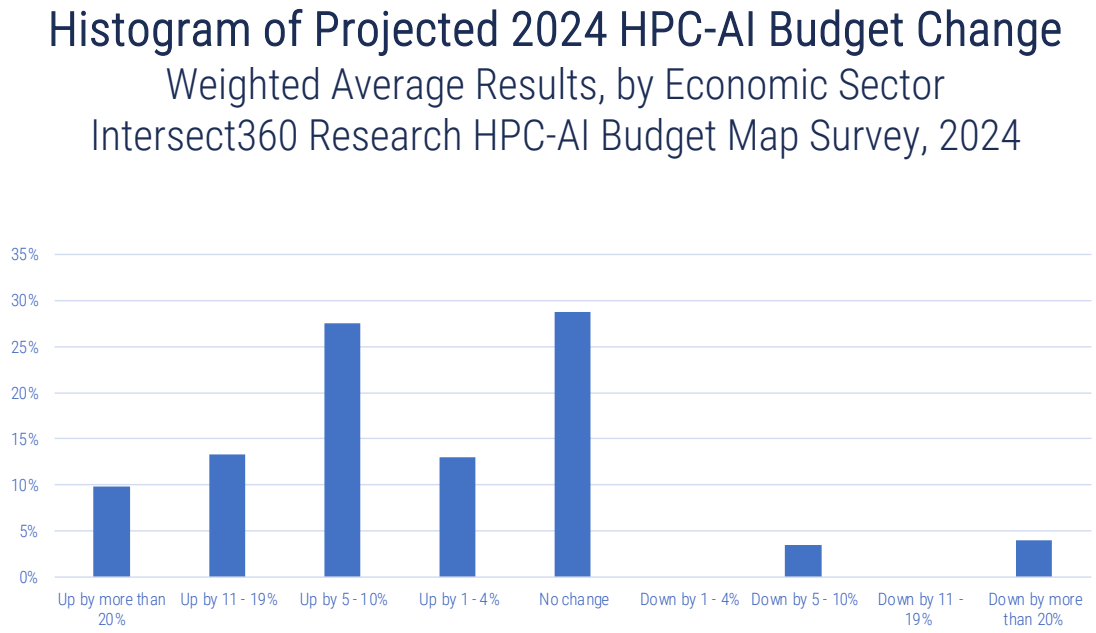
- Midyear check on 2024 revenue for major suppliers, including HPE, Dell, Supermicro, Lenovo, Nvidia, Intel, AMD, …
- Most are trending to 75% to 150% growth
- HPE and Dell are usually bellwethers for on-premises HPC-AI – both are recognizing major hyperscale AI revenue this year



©2024 Intersect360 Research

HPC-AI Budget Expectation Data Roll-Up

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



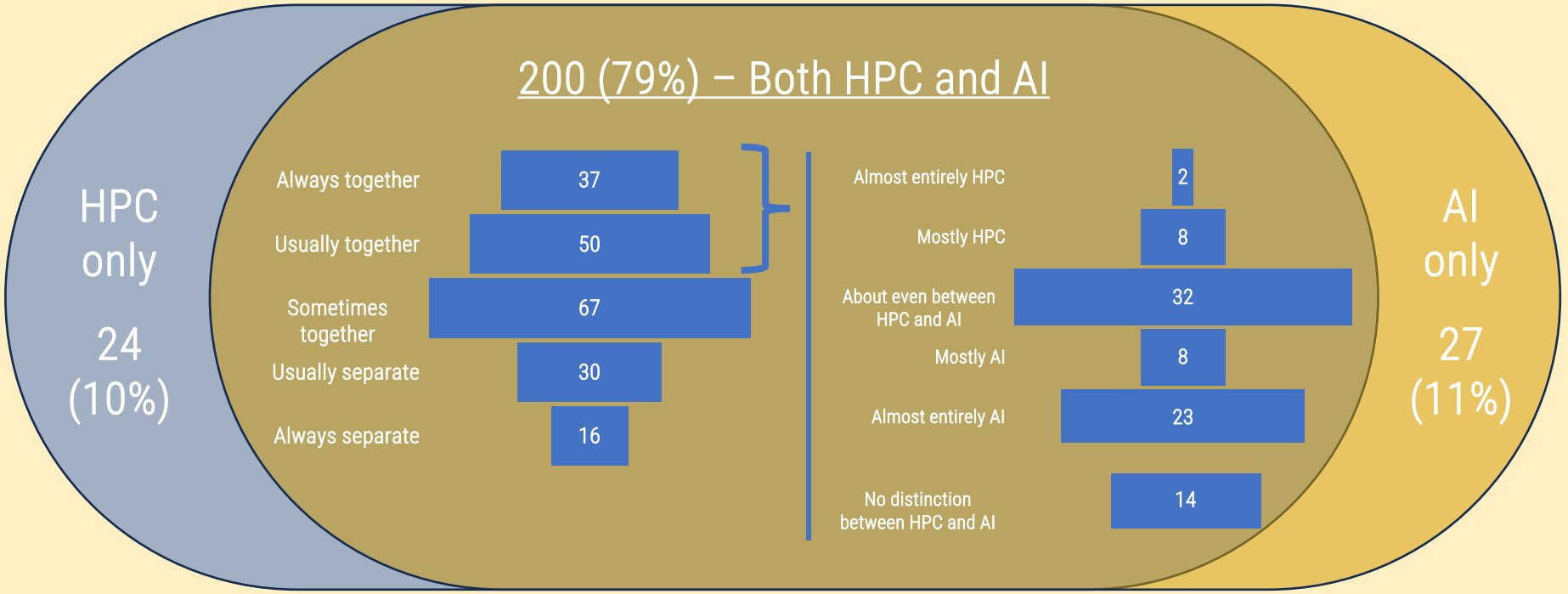
©2024 Intersect360 Research

HPC-AI Budgets among U.S. Enterprise

264 total contacted

251 (95%) had HPC or AI budgets

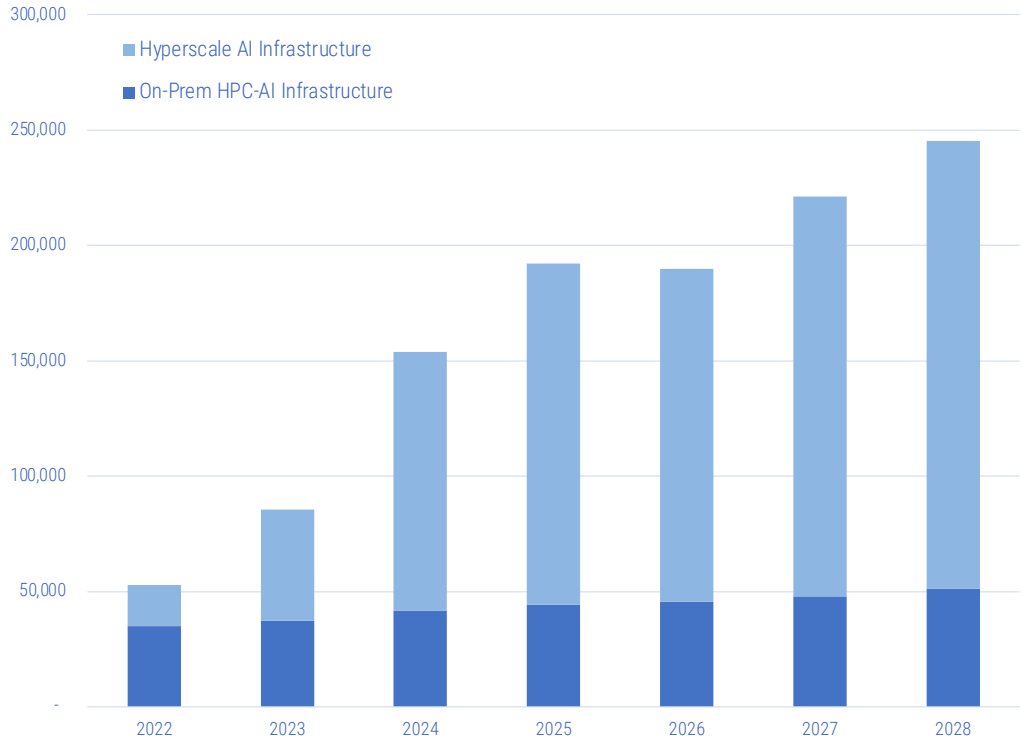
13 (5%) neither HPC nor AI



©2024 Intersect360 Research

Revised HPC-AI Infrastructure Forecast ($M)

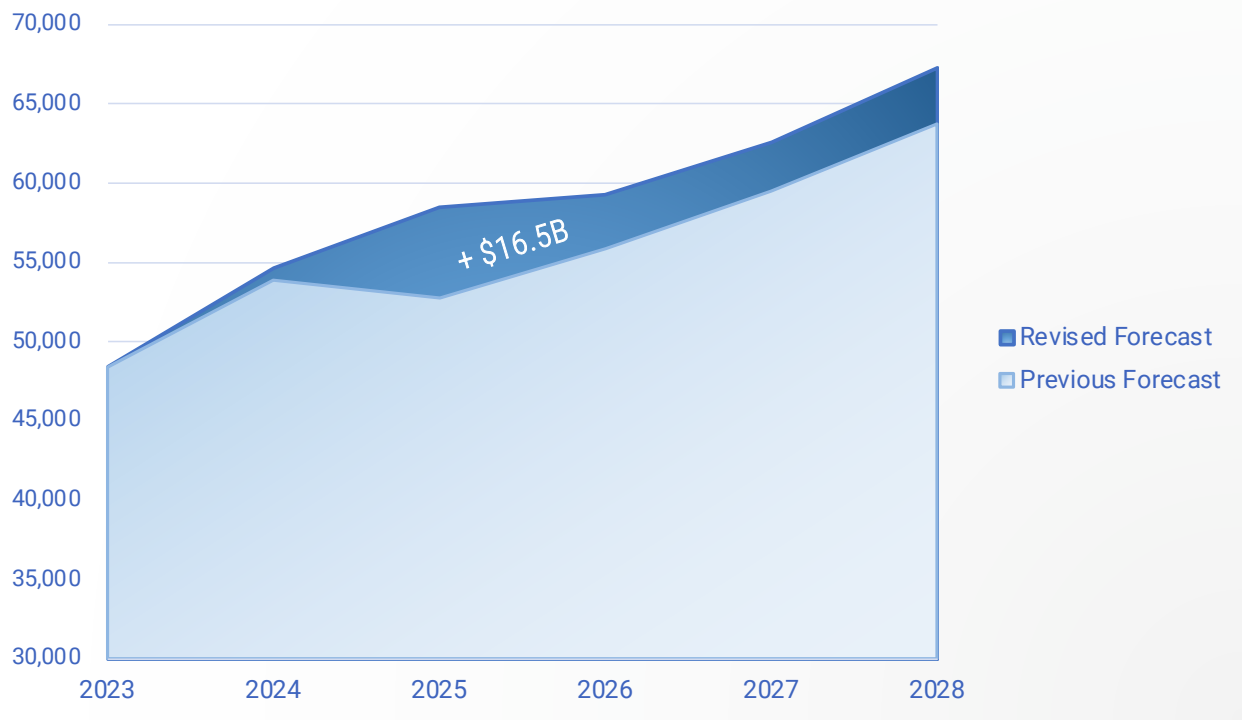
- Hyperscale AI has second-straight year of triple-digit growth
- Hyperscale AI in 2024 is more than 6x where it was in 2022
- Hyperscale AI segment will near $200 billion in 2028
- On-premises HPC-AI infrastructure now forecast to grow 11.8% in 2024 (was 11.0% in May 2024 forecast)
- Increase in on-premises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



©2024 Intersect360 Research

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

- Slight increase to outlook for this year
- Biggest difference is in 2025 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)



©2024 Intersect360 Research

Revised HPC-AI 2024-28 Market Forecast

- Growth in on-premises enterprise AI pushes market higher than previous forecast
- Slowdown occurs in 2026 (delayed from 2025) as business models catch up to enthusiasm
- The AI boom is net additive to the market, with long-term higher growth rate
- Total of $16.5B in added revenue through 2028 compared to previous forecast
- Non-hyperscale HPC-AI market five-year CAGR: 6.8%

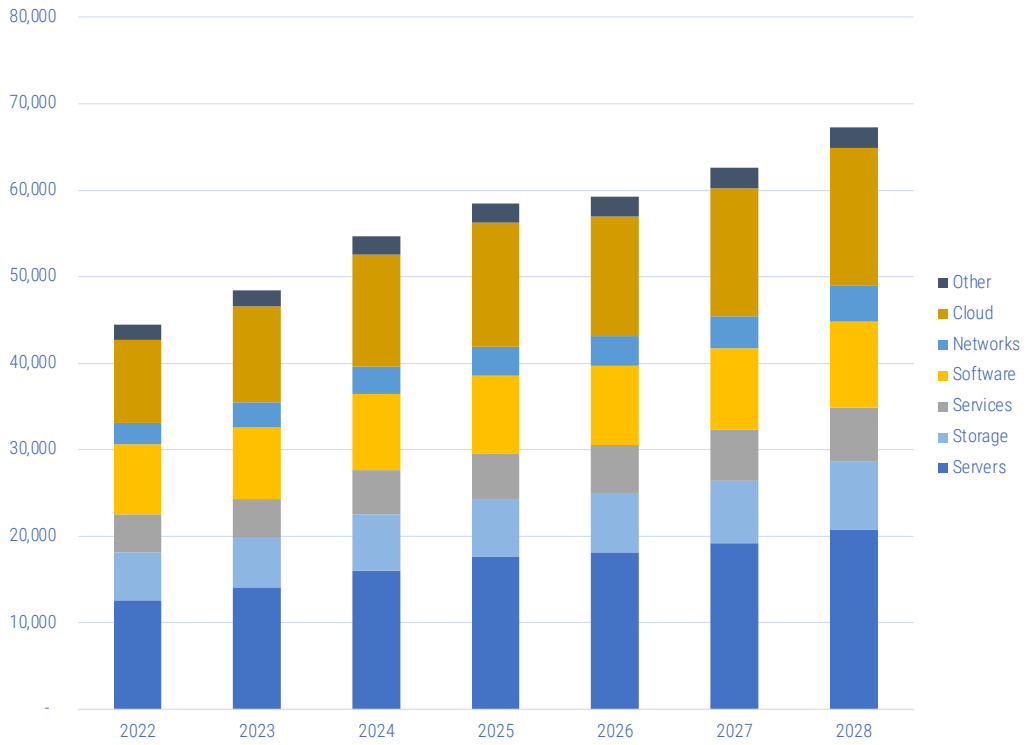
©2024 Intersect360 Research

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

This is the primary segmentation analyzed for the midyear forecast update.

Other segmentations are back-calculated using previously forecast allocations.

See May 2023 forecast for full methodology details.



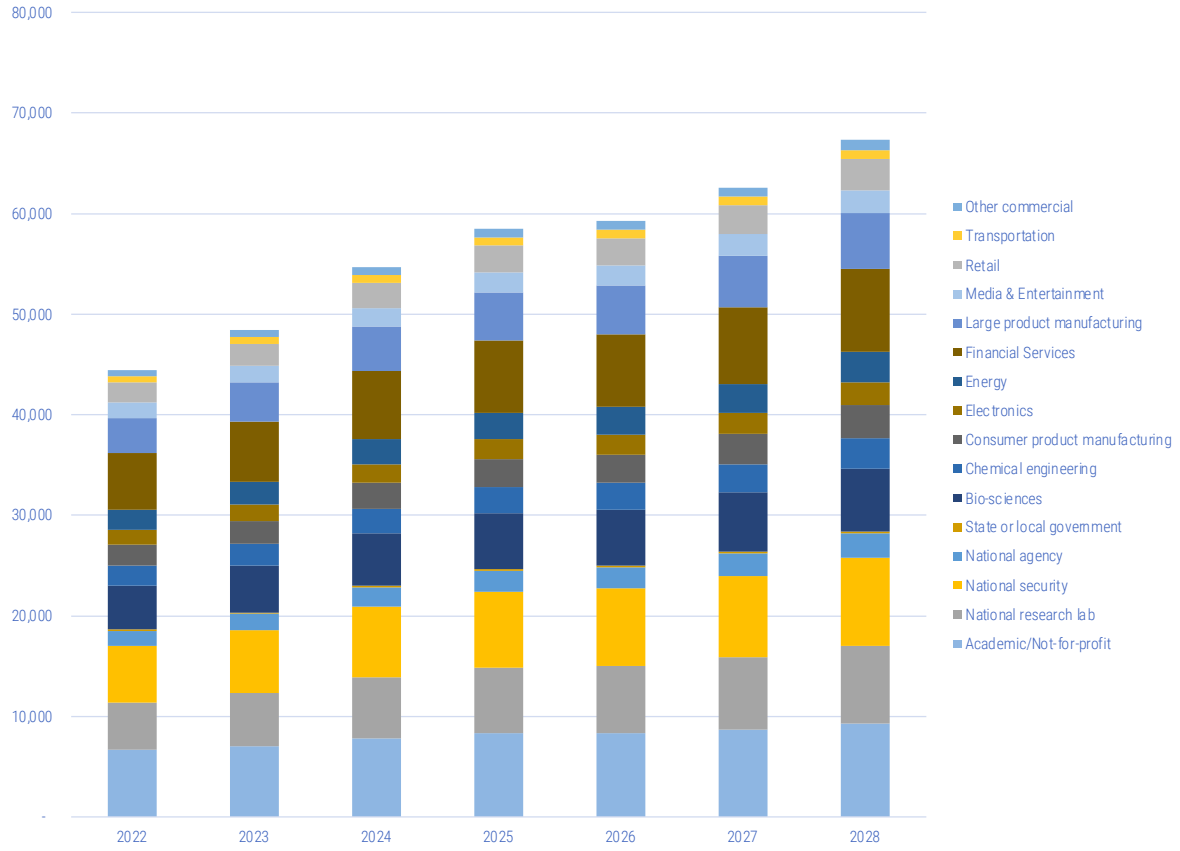
©2024 Intersect360 Research

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

HPC-AI MARKET, EXCLUDING HYPERSCALE	2022	2023	2024	2025	2026	2027	2028	5yr CAGR 2023-28
Servers	12,601	14,065	15,993	17,637	18,166	19,165	20,794	8.1%
Storage	5,494	5,725	6,568	6,590	6,857	7,269	7,803	6.4%
Services	4,464	4,480	5,059	5,323	5,509	5,836	6,312	7.1%
Software	8,086	8,350	8,819	9,042	9,196	9,445	9,941	3.5%
Networks	2,487	2,823	3,201	3,307	3,421	3,698	4,064	7.6%
Cloud	9,577	11,109	12,946	14,374	13,845	14,797	15,992	7.6%
Other	1,722	1,874	2,063	2,190	2,258	2,381	2,402	5.1%
Total	**44,431**	**48,425**	**54,647**	**58,462**	**59,252**	**62,591**	**67,308**	**6.8%**

©2024 Intersect360 Research

Revised HPC-AI Forecast: Vertical Markets ($M)



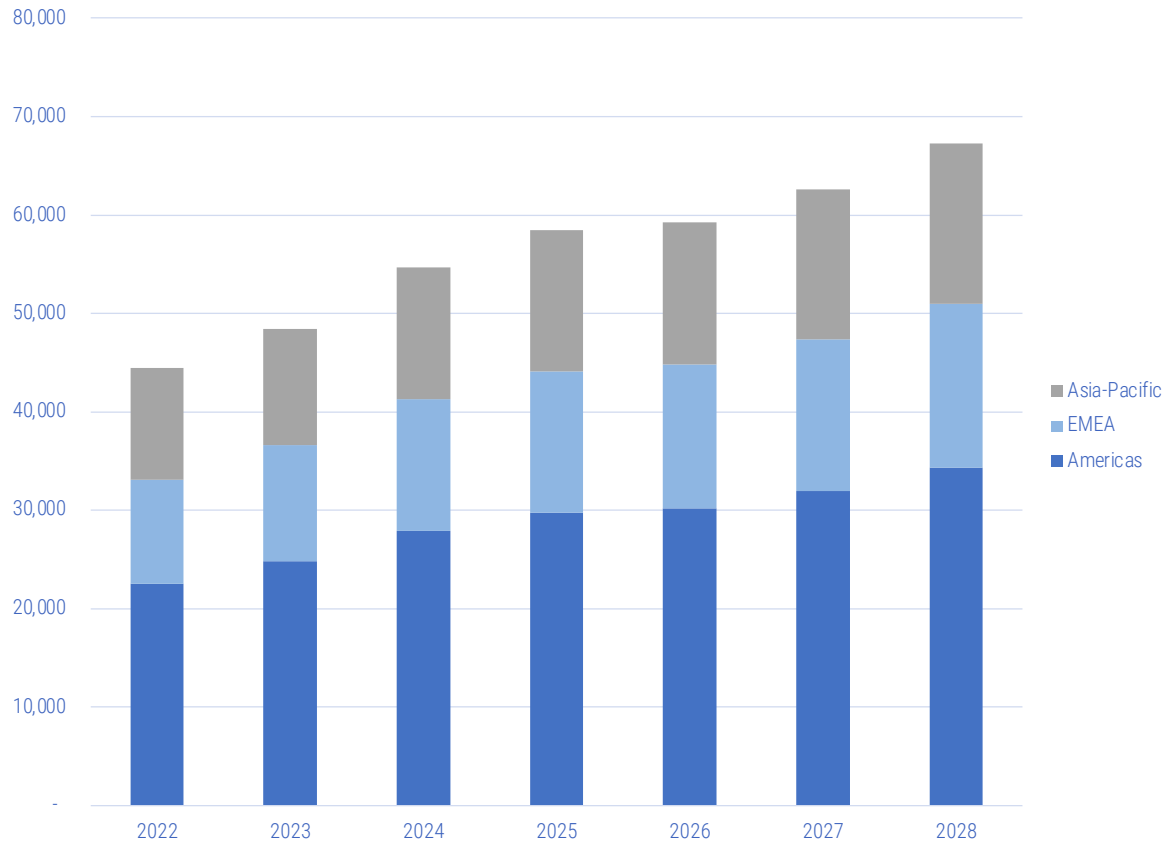
©2024 Intersect360 Research

Revised HPC-AI Forecast: Vertical Markets ($M)

HPC-AI MARKET, EXCLUDING HYPERSCALE	2022	2023	2024	2025	2026	2027	2028	5yr CAGR 2023-28
Academic/Not-for-profit	6,665	7,022	7,832	8,290	8,313	8,688	9,241	5.6%
Government:								
National research lab	4,713	5,277	6,035	6,532	6,696	7,154	7,780	8.1%
National security	5,660	6,246	7,070	7,572	7,684	8,126	8,749	7.0%
National agency	1,482	1,638	1,885	2,034	2,080	2,217	2,405	8.0%
State or local government	141	156	179	193	197	209	227	7.8%
Commercial:								
Bio-sciences	4,319	4,651	5,211	5,535	5,570	5,843	6,238	6.0%
Chemical engineering	1,993	2,165	2,448	2,619	2,653	2,802	3,013	6.8%
Consumer product manufacturing	2,085	2,300	2,599	2,795	2,847	3,022	3,267	7.3%
Electronics	1,494	1,635	1,837	1,967	1,995	2,109	2,269	6.8%
Energy	2,022	2,216	2,496	2,674	2,715	2,872	3,093	6.9%
Financial Services	5,572	6,033	6,764	7,201	7,263	7,634	8,169	6.2%
Large product manufacturing	3,505	3,862	4,394	4,734	4,832	5,141	5,567	7.6%
Media & Entertainment	1,536	1,666	1,877	2,005	2,029	2,140	2,298	6.7%
Retail	2,002	2,194	2,474	2,651	2,691	2,846	3,065	6.9%
Transportation	635	688	778	831	841	887	953	6.7%
Other commercial	607	677	768	828	846	900	975	7.6%
Total	**44,431**	**48,425**	**54,647**	**58,462**	**59,252**	**62,591**	**67,308**	**6.8%**

©2024 Intersect360 Research

Revised HPC-AI Forecast: Regions ($M)



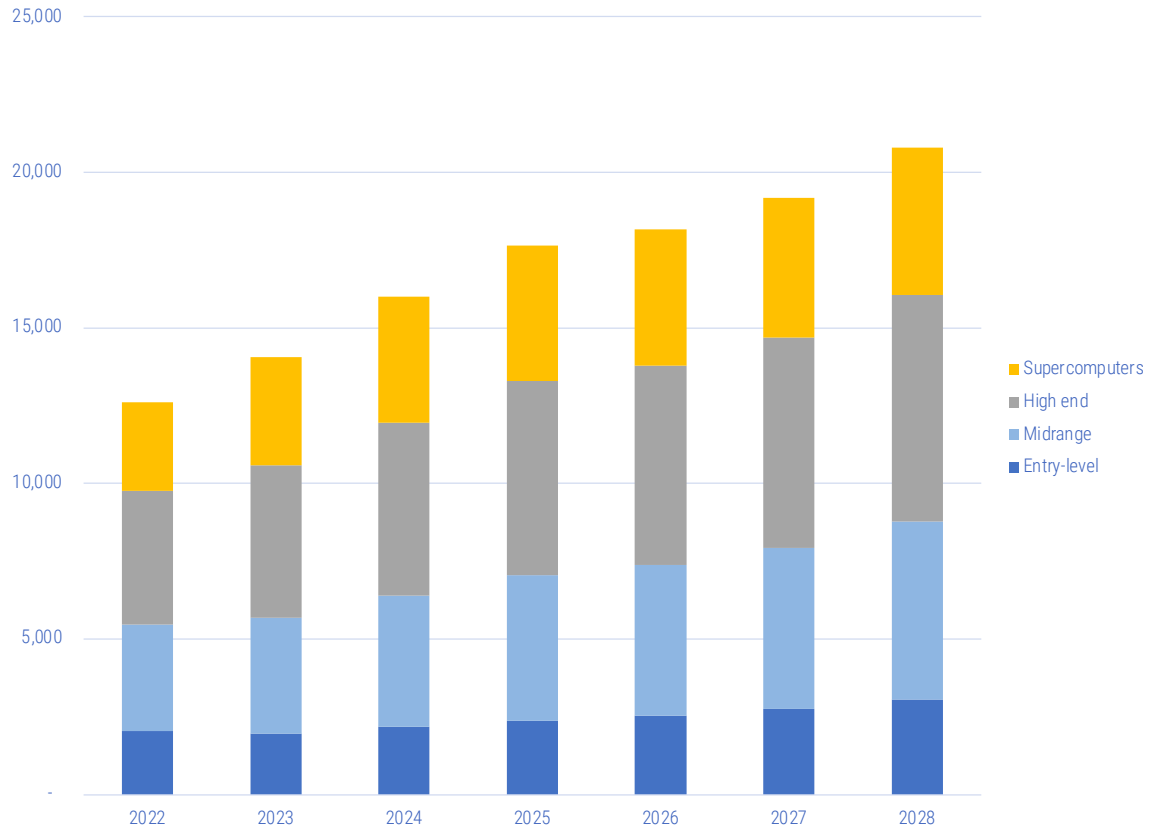
©2024 Intersect360 Research

Revised HPC-AI Forecast: Regions ($M)

HPC-AI MARKET, EXCLUDING HYPERSCALE	2022	2023	2024	2025	2026	2027	2028	5yr CAGR 2023-28
Americas	22,518	24,842	27,916	29,753	30,192	31,932	34,381	6.7%
EMEA	10,627	11,767	13,388	14,363	14,583	15,433	16,625	7.2%
Asia-Pacific	11,286	11,816	13,344	14,346	14,477	15,226	16,302	6.6%
Total	**44,431**	**48,425**	**54,647**	**58,462**	**59,252**	**62,591**	**67,308**	**6.8%**

©2024 Intersect360 Research

Revised HPC-AI Forecast: On-Premises HPC-AI Server Classes ($M)



©2024 Intersect360 Research

Revised HPC-AI Forecast: On-Premises HPC-AI Server Classes ($M)

ON-PREM HPC-AI SERVERS, EXCLUDING HYPERSCALE	2022	2023	2024	2025	2026	2027	2028	5yr CAGR 2023-28
Supercomputers	2,832	3,488	4,030	4,356	4,369	4,485	4,731	6.3%
High end	4,305	4,895	5,566	6,231	6,404	6,743	7,301	8.3%
Midrange	3,430	3,727	4,222	4,665	4,865	5,196	5,706	8.9%
Entry-level	2,033	1,955	2,175	2,385	2,528	2,742	3,057	9.4%
Total	**12,601**	**14,065**	**15,993**	**17,637**	**18,166**	**19,165**	**20,794**	**8.1%**

©2024 Intersect360 Research

Conclusions

- Generative AI is having a major effect on the HPC-AI market, adding significant growth.
- Hyperscale computing has become dominant and will loom over the HPC-AI market.
- The long-term market outlook has increased; there is a stable need for HPC technologies.
- By 2026, there will be an emphasis on business models for enterprise AI.
- There is significant risk in any forecast. The hyperscale market is in high growth but is unstable. Changes in global geopolitical relationships could dramatically alter the market.
- Refer to May 2024 forecast for full methodology. Contact Intersect360 Research for inquiries.

©2024 Intersect360 Research

HALO Worldwide Survey Results

November 2024

Paul Muzio, Steve Conway

Thank You to Our HALO Advisory Committee Members

HALO
HPC-AI LEADERSHIP ORGANIZATION

HALO EMEA


Jean-Philippe Nominé
CEA
(Chair)


Rosa Badia
Barcelona Supercomputing Ctr


Henri Calandra
TotalEnergies


Christine Kitchen
ECMWF


Hatem Ltaief
KAUST


Marek Magryś
Cyfronet Poland


Stephan Schenk
BASF


Happy Sithole
CHPC South Africa


Riccardo Testi
Piaggio

Gerd Büttner, Airbus

Tim Stitt, Roche

HALO Americas


Dan Stanzione
TACC
(Chair)


Carlos J. Barros H.
U. Santander


Rupak Biswas
NASA Ames Research Center


Frank Indiviglio
NOAA


Brock Kahanyshyn
Digital Research Alliance Canada


Elizabeth L'Heureux
BP


William Edsall
Dow Chemical


Gina Tourassi
ORNL


Marcus Weber
Caterpillar Inc.

Mike Costantino, NYU Langone

Unnamed, Financial Institution

HALO Asia-Pacific


Mark Stickells
PAWSEY (Chair)


Satoshi Matsuoka
RIKEN

Issues Facing the HPC-AI Industry

Insights from HALO Advisory Committee interviews

HALO
HPC-AI LEADERSHIP ORGANIZATION

Full Report Now Available

HALO
HPC-AI LEADERSHIP ORGANIZATION

Intersect360 RESEARCH

ISSUES FACING THE HPC-AI INDUSTRY: INSIGHTS FROM THE ADVISORY COMMITTEES OF THE HPC-AI LEADERSHIP ORGANIZATION (HALO)

ADDISON SNELL STEVE CONWAY

PAUL MUZIO KEVIN JACKSON

OCTOBER 2024

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

© 2024 Intersect360 Research. Information from this report may not be distributed in any form without permission from Intersect360 Research.

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Infrastructure Design:** Challenges in creating optimal HPC-AI environments, balancing homogenous and heterogeneous systems, and integrating diverse processor types.
- **Human Resources:** The shortage of skilled personnel in computational sciences and HPC-AI system management, regional disparities, and competition for talent.
- **Porting and Developing Applications:** Issues related to application portability across different systems and the challenges in porting and validating applications.
- **Accuracy / Reproducibility of Results:** Concerns about result consistency across diverse chip technologies and verification challenges in different hardware and software configurations.
- **Processor Suitability and Market Issues:** A wide variety of processor requirements for different applications, GPU demand and pricing implications, and concerns about processor supply and development.
- **AI/LLM Training and Use:** Challenges in data availability, ownership, legal restrictions, and model creation and validation.
- **System Software Stacks:** Issues related to HPC Nationalism, the integration of AI and traditional HPC support, and the need for improved schedulers and file systems.
- **Sustainability:** Concerns about power consumption in HPC-AI facilities and its impact on infrastructure and chip design.

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Infrastructure Design:** Challenges in creating optimal HPC-AI environments, balancing homogeneous and heterogeneous systems, and integrating diverse processor types.
  - On prem, colo, hyperscalers, cloud…
  - Remaining influence of Top500?
  - Tension between HPC and AI requirements

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Human Resources:** The shortage of skilled personnel in computational sciences and HPC-AI system management, regional disparities, and competition for talent.
  - US-Europe (EMEA?) less dynamic than Asia-Pacific
  - Tension on job market / salaries (AI for the better or for the worse?)

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Porting and Developing Applications:** Issues related to application portability across different systems and the challenges in porting and validating applications.
  - Specialization vs. portability
  - Promises of containers/virtualization (w/o sacrificing perf?)

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Accuracy / Reproducibility of Results:** Concerns about result consistency across diverse chip technologies and verification challenges in different hardware and software configurations.
  - Wide choice of compute units, good for specialization, less good for verification/validation
  - Industrial codes may not map well to GPUs or other accelerators

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **Processor Suitability and Market Issues:** A wide variety of processor requirements for different applications, GPU demand and pricing implications, and concerns about processor supply and development.
  - Again, AI for the better and for the worse! Chips features, pricing
  - Market/technology trend distortion – Top500 extra bias?
  - Possible trouble in the supply chain (GPU quasi monopoly, supply limitations, export control…)

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **AI/LLM Training and Use:** Challenges in data availability, ownership, legal restrictions, and model creation and validation.
  - Regional regulatory discrepancies
  - Are LLMs good or bad for language/culture diversity?
  - Urgent need for smaller models

HALO
HPC-AI LEADERSHIP ORGANIZATION

Issues Facing HPC-AI Identified in HALO Advisory Committee Interviews

- **System Software Stacks:** Issues related to HPC Nationalism, the integration of AI and traditional HPC support, and the need for improved schedulers and file systems.
  - Risk of regional silos (chips, software…)
  - HPC-AI converged software stacks?
  - Schedulers and filesystems: for larger (Exascale HPC) AND more diverse (AI) workloads and data flows

HALO
HPC-AI LEADERSHIP ORGANIZATION

SC24 Preview

HALO

HPC-AI LEADERSHIP ORGANIZATION

SC24 Reception

intersect360.com/sc24

Intersect360 Research SC24 Reception

Monday, Nov 18 2:00 - 4:00 p.m.

Max Lager's Wood-Fired Grill & Brewery

320 Peachtree St NE, Atlanta

HALO
HPC-AI LEADERSHIP ORGANIZATION