Special Analysis

GPU and Accelerator Growth in HPC

Earl Joseph and Alex Norton
December 2020

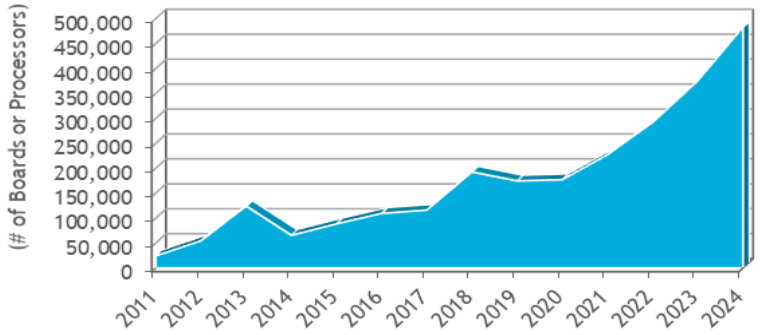
HYPERION RESEARCH OPINION

Hyperion Research has followed the growth of accelerators in HPC for more than 10 years. Accelerators had an initially slow HPC adoption curve but have recently experienced amplified growth. Within the HPC accelerator market, GPUs, particularly those from Nvidia, have dominated. Although originally designed as gaming processors for graphics rendering, GPUs were found to be well suited to a range of HPC applications, plus critical underlying AI functions, such as matrix multiplication, driving impressive performance gains. As a result, many HPC sites across a broad range of verticals are now using accelerators to speed up a larger portion of their increasingly AI-based workloads.

The historical growth rate in installing GPUs and accelerators is over 48% CAGR and is projected to grow by over 22% a year for the next five years. This represents more than triple the growth rate for accelerators compared with HPC servers over the same forecast period. As AI workloads become an increasingly important part of overall HPC workload, both enabling new innovative techniques to solve new problems and to supplement traditional modelling and simulation workloads, GPU usage will likely continue to have strong growth. Figure 1 below shows Hyperion Research's projection for accelerators in the HPC market.

FIGURE 1

Technical Computer Market: GPUs/Accelerators/FPGA and Other Types of Accelerators



Source: Hyperion Research, 2020

GPU AND ACCELERATOR ADOPTION IN HPC

The use of accelerators, and specifically GPUs, has grown rapidly over the last decade, shown in Table 1 below. Many HPC sites and workloads are increasingly using accelerators in system procurements in a response to the growing importance of HPC-based of AI workloads. However, GPUs are not the only accelerators available or used in the HPC market, although they dominate the majority of installations. FPGA (field-programmable gate arrays) have shown promise and adoption in certain sectors of the market, like bio-sciences and finance, although they can be more complex to integrate and program as compared with GPUs. A major provider of FPGA technology for HPC is Xilinx, although other companies such as Intel, through their acquisition of Altera, provide FPGAs to the HPC, and broader IT, market. One of the key differentiators of FPGAs is their ability to be reprogrammed for different applications, providing users with more flexibility in the utilization of such technology.

TABLE 1

Technical Computer Market: GPUs/Accelerators/FPGA and Other Types of Accelerators

	2011	2012	2013	2014	2015	2016	2017	2018	2019	Historical CAGR '11-'19
Total Attached Processors or Boards	23,902	54,251	123,678	65,092	88,123	109,166	115,637	192,162	174,190	48.8%
Yearly (YOY) Growth		127.0%	128.0%	-47.4%	35.4%	23.9%	5.9%	66.2%	-9.4%	

Note: 2013 and 2018 had a few very large systems installed.

Source: Hyperion Research, 2020

Table 2, below, shows the four major competitive segments tracked in the HPC market. The Supercomputer segment, which represents the most expensive machines in the world, priced at $500,000 and above, exhibits the highest use of accelerators, due to the larger scale of workloads typical for these systems as well as a greater emphasis on performance over price than in other HPC price bands. The two major peaks seen in the historical data for accelerator use, 2013 and 2018, represent major individual purchases of large supercomputers relying heavily on accelerator technology. The most recent of those, the Summit and Sierra systems installed at Oak Ridge National Laboratory and Lawrence Livermore National Laboratory, respectively, both rely heavily on Nvidia GPU technology.

TABLE 2

GPU/Accelerator Historical Market by Competitive Segments

	2011	2012	2013	2014	2015	2016	2017	2018	2019	Historical CAGR '11-'19
Supercomputer	13,165	37,023	98,816	34,517	42,368	52,802	46,934	111,301	76,587	42.2%
Divisional	4,192	5,876	8,209	11,928	18,695	25,166	33,813	44,155	52,635	65.9%
Departmental	3,286	6,906	8,512	10,319	17,479	20,613	21,404	23,374	30,079	55.7%
Workgroup	3,259	4,446	8,141	8,328	9,581	10,583	13,485	13,332	14,889	35.5%
Total	23,902	54,251	123,678	65,092	88,123	109,166	115,637	192,162	174,190	48.8%

Source: Hyperion Research, 2020

Inhibitors to GPU/Accelerator Adoption

Despite their rapid growth, accelerators still face some barriers to adoption. Many workloads were not designed to exploit the capabilities of accelerated computing and often need to be refactored and modernized to run on CPU-accelerator platforms a complex, and often unrewarding task. In addition, some machines have been designed, tuned, and optimized for a specific architecture that may not be well suited to take advantage of the particular performance capabilities of available accelerators. Furthermore, accelerators can be a costly option for an overall system design, requiring additional and expensive power and cooling capabilities.

FUTURE ACCELERATOR ADOPTION

Accelerators are anticipated to grow in deployment over the next five years, as can be seen in Tables 3 and 4 below. An interesting note is that the three major DOE systems, which are slated to be installed in the 2021-2023 time period, do not sway the yearly growth figures as much as in years past. The large DOE systems and others like it around the world will be responsible for major accelerator deployments over the period. In addition, the broader market beyond the top one percent of the highest-end machines is also increasing their use of GPUs.

In the forecast period, GPUs are anticipated to contribute a large majority of accelerators installed in HPC, namely from Nvidia, but also from AMD and Intel. AMD's Radeon Instinct GPU is slated to be the accelerator of choice for two of the CORAL-2 machines to be installed in 2021 and 2022, and Intel's Xe GPU is to be deployed in the Aurora machine. While the middle two segments of the market, Departmental and Divisional, are increasing their adoption, the Supercomputer segment shows the highest growth of accelerator adoption. The larger budgets for Supercomputers allow users to incorporate more heterogeneity into their system (e.g. more accelerators and different types of accelerators), addressing a broad set of workload requirements. (See Table 4 below)

TABLE 3

Technical Computer Market Forecast: GPUs/Accelerators/FPGA and Other Types of Accelerators

	2019	2020	2021	2022	2023	2024	Forecasted CAGR '19-'24
Total Attached Processors or Boards	174,190	176,647	226,760	291,559	375,002	482,466	22.6%
Yearly Growth	-9.4%	1.4%	28.4%	28.6%	28.6%	28.7%	

Note: 2013 and 2018 had a few very large systems installed.

Source: Hyperion Research, 2020

TABLE 4

GPU/Accelerator Forecast by Competitive Segments

	2019	2020	2021	2022	2023	2024	Forecasted CAGR '19-'24
Supercomputer	76,587	84,093	115,776	165,850	231,878	320,824	33.2%
Divisional	52,635	50,003	64,159	73,340	85,970	99,744	13.6%
Departmental	30,079	29,096	32,971	38,105	42,468	46,776	9.2%
Workgroup	14,889	13,455	13,853	14,264	14,686	15,121	0.3%
Total	174,190	176,647	226,760	291,559	375,002	482,466	22.6%

Source: Hyperion Research, 2020

GPU and Accelerator Use by Verticals

Table 5, below, highlights the forecast adoption of accelerators by the 13 HPC verticals tracked by Hyperion Research. Government Labs, today and continuously through the forecast period, represent the highest adoption of accelerators by four times more than the second largest vertical, Manufacturing. Projecting out five years, six verticals are anticipated to exhibit five year CAGRs of more than 20%, with Bio-Sciences showing the highest growth at more than 30% CAGR. These verticals are both adopting and integrating AI for many of their standard workloads, but also working to modernize code to take advantage of the performance gains possible with accelerated computing.

TABLE 5

GPU/Accelerators Forecast by Industry/Application Segments (In Number of Boards or Processors)

	2019	2020	2021	2022	2023	2024	Forecasted CAGR '19-'24
Bio-Sciences	5,176	5,588	7,635	10,450	14,308	19,596	30.5%
CAE	23,632	24,556	32,300	43,293	56,935	73,562	25.5%
Chemical Engineering	2,126	2,177	2,864	3,293	3,825	5,019	18.7%
DCC & Distribution	3,981	4,025	5,228	6,312	7,302	9,460	18.9%
Economics/Financial	5,142	5,189	6,728	8,234	9,736	12,589	19.6%
EDA	10,654	10,847	14,192	16,467	20,167	26,314	19.8%
Geosciences	4,535	4,323	5,216	6,304	7,622	9,218	15.2%
Mechanical Design	69	64	74	77	81	84	4.0%
Defense	9,327	9,420	12,227	15,369	19,124	24,481	21.3%
Government Lab	81,647	81,091	103,003	131,046	166,782	212,323	21.1%
University/Academic	16,526	16,962	22,037	28,677	37,331	48,610	24.1%
Weather	2,787	2,911	3,804	4,980	6,521	8,451	24.8%
Other	8,588	9,496	11,450	17,054	25,269	32,760	30.7%
Total	174,190	176,647	226,760	291,559	375,002	482,466	22.6%

Source: Hyperion Research, 2020

FUTURE OUTLOOK

The accelerator market will be a high growth market for the foreseeable future, as utilization and integration of new HPC and AI applications and the desire to modernize a wide base of existing code for many HPC sites continues to grow. A majority of accelerators deployed over the next five years will likely be GPUs, building on their existing market leadership and established mind share. Hyperion Research is anticipating that FPGAs and ASICs will increase their market presence over the next five years, as small and large companies alike seek to provide users with alternatives to main-stream GPU options. Many users will also take advantage of the cloud to experiment with and use accelerators ahead of future procurements or for subsets of their workloads that do not warrant an on-premise deployment of accelerators, whether from a system design perspective or a cost analysis.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
