HPC and AI Processors: An Update

May 2021

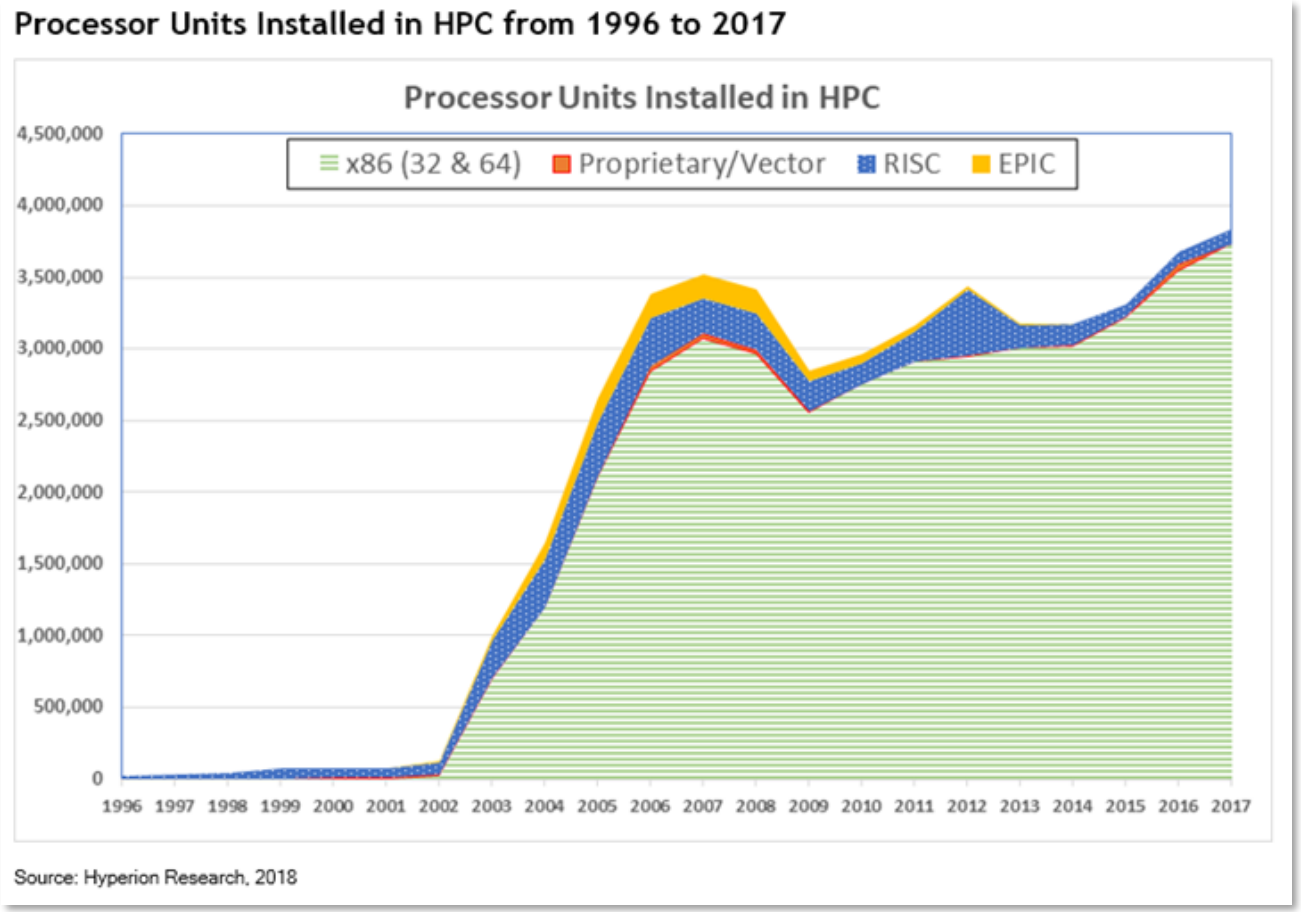
Alex Norton

www.HyperionResearch.com
www.hpcuserforum.com

x86 Dominance

x86 processors have remained the market standard for nearly two decades

- **x86 processors have consisted of at least 90% of the processors installed in HPC since 2006, just four years after their emergence in the market**
  - Many applications today are written first for x86 systems
  - Although many architectures are trying, none yet have supplanted x86 as the HPC processor of choice



x86 Vendor Competition

Intel has held command of the HPC x86 market for many years now

- **Recently, AMD has brought a product to the market that poses a strong challenge to the Intel market share**
- **AMD's EPYC line has gained traction, not only through its selection as the key processor for the CORAL-2 DOE procurements, but also through strong price/performance and raw performance benchmarks**
  - Despite this surge, Intel remains the dominant x86 supplier
- **The future ecosystem of HPC processors looks to be diversifying, with the influx of x86 vendor alternatives as well as Arm and other processor architectures and types**
- **The future designs of HPC datacenters are anticipated to be heterogeneous by nature, with technologies in systems to handle the most diverse workloads**

Where Does Arm Factor In?

Arm presents an interesting alternative to x86

- **Arm technology has been thrown into the limelight once again with the Fugaku machine at RIKEN**
  - Fugaku is based on the A64fx chip, an Arm-based processor developed by Fujitsu and RIKEN
  - The A64fx chip demonstrates not only the capability of Arm from the perspective of flexibility but also in co-design between application and processor experts
- **Arm has experienced other major deployments, including multiple systems in the UK as well as part of the HPE Moonshot program with the US DOE**
  - Porting codes to Arm can be difficult and time consuming, but for some applications it has shown positive performance

Arm Forecast

Arm processors growing in adoption by HPC users

(Processor Counts)	2019	2020	2021	2022	2023	2024	CAGR '19-'24
All HPC Processors	4,086,750	3,703,232	4,100,843	4,836,508	5,447,354	5,713,783	6.9%
ARM Processors	44,455	161,502	329,765	321,354	390,746	447,028	58.7%
ARM as % of all HPC processors	1.1%	4.4%	8.0%	6.6%	7.2%	7.8%	48.4%
Source: Hyperion Research, 2021							

- **Arm is rapidly becoming a more important part of the HPC processor ecosystem, fueled by the deployment of the Fugaku machine (the rapid growth from 2020 to 2021 above)**
  - Although Arm is anticipated to grow at 48.4% CAGR for the next four years, it will represent just under 8% of the total processors shipped in HPC in 2024

Purpose-Built Processors by End Users

End users are increasingly building custom processors for their specific needs

- **As workloads have become increasingly diverse in the HPC space, especially with the injection of AI, end users have started to design chips to their specific needs**
- **Many of these chips are based on Arm technology, including the Fujitsu chip mentioned earlier**
- **Other notable purpose-built processors include:**
  - AWS's push towards purpose-built Arm processors. AWS developed the Graviton2 chip, a custom Arm processor on their cloud platform, as well as the Trainium and Inferentia chips, which target the AI market specifically
  - Apple's M1 chip made a splash in the general IT market, due primarily to the way it incorporates different accelerator cores to address specific workloads or actions performed by the chip
  - Google's TPU has provided strong competition to the GPU since development, but was originally developed specifically for Google's needs
  - Other noted efforts include Facebook's push to have custom processors, as well as China and Europe's national efforts for purpose-built processor for AI, HPC, and other application areas

Risks in HPC

A shrinking number of high-capability fab companies worldwide means a battle for producing 7nm, 5nm, and 3nm processors en masse

- **And then there were three: TSMC (Taiwan), Samsung (S Korea) and Intel (US)**

Process node (nm)	180	130	90	65	45/40	32/28	22/20	16/14	10/7	5	3
Number of semicondutor manufacturers working at each process node											
US	24	18	11	8	4	4	4	4	1	1	1
South Korea	4	4	3	2	2	2	2	2	2	1	1
Taiwan	9	9	6	6	6	6	5	3	1	1	1
Japan	18	10	7	6	5	1	1	1			
China	19	18	16	13	8	6	3	1	1		
Other	20	13	5	1	1	1	1				
Total	*94*	*72*	*48*	*36*	*26*	*20*	*16*	*11*	*5*	*3*	*3*

Note: Some companies in the above table have fabrication facilities located in countries outside of where they are headquarted but have been included in country totals. The table also does not distinguish between producers of different types of semiconductors, such as CPU/GPU, application-specific semiconductors, and memory, each of which is driven by different market requirements around feature size.

Source: The Eurasia Group, 2020, https://www.eurasiagroup.net/live-post/geopolitics-semiconductors

-N.B. Apple's M1 chip uses about 25 percent of TSMC's 5-nm production capacity

Accelerators and Co-Processors

NVIDIA GPU: The Market Standard

As with Intel x86, NVIDIA GPUs have dominated the accelerator market for years

- **NVIDIA GPUs burst onto the scene as graphics cards for gaming, but soon it becomes clear that for certain AI applications, the GPU has high performance advantages over CPU-only training**
  - GPUs have shown strong adoption in the HPC ecosystem, especially as AI has been injected into many HPC sites
  - NVIDIA's CUDA has been the strongest part of the GPU family, enabling users to take more advantage of the GPU platform through the programming and library side of applications
- **There are, however, drawbacks to GPUs:**
  - Cost – GPUs can be very expensive to deploy in HPC datacenters
  - Power-efficiency – GPUs require more power to run, increasing their operating cost
  - "General" applicability – not all HPC applications can benefit today from accelerated computing, leading to the possibility of GPUs being under-utilized in a system

General Accelerator Competition

A profusion of devices to meet diversifying workloads

- **NVIDIA GPUs are encountering competition in the marketplace for general use accelerators (i.e., accelerators for more than just AI applications)**
  - AMD has already released their GPU (Radeon), which is slated to be the accelerator of choice in two exascale machines in the US
  - Intel is adding a GPU to their portfolio of accelerators, which already includes FPGAs as well as other, more unique accelerators
  - Xilinx's FPGA is starting to gain more widespread appeal for general accelerators, although more challenging to program than GPUs
- **General accelerators will dominate the market as being able to accelerate existing workloads of end users, as well as handle the needs of emergent and novel applications, like new ML or DL applications, being introduced in the HPC space**
  - The other key factor is updating codes to be able to take advantage of deployed accelerators, something CUDA from NVIDIA has established as a necessary practice

Novel AI Accelerators

Dedicated AI hardware entering market to address specific needs of users

- **In addition to the rise of more general accelerators in the HPC and AI spaces, there has also been an emergence of specialized processors and accelerators targeted at specific segments of the AI market**
  - Some of these processors are designed specifically to handle training workloads in a datacenter, while others are designed to be low power, edge processors for inference workloads
  - Processor companies are also focused on specific workloads, like image or video recognition, or NLP, while others focus on more general AI performance
- **The market for these novel processors and accelerators is still nascent but filled with energy and a large amount of financial investment**
  - The companies involved are not necessarily always startups. Intel has a neuromorphic chip in development that is not available yet, through Intel Labs

Highlighted Deployments of AI Accelerators

Novel AI accelerators have made some headlines lately

- **This list is not exhaustive, but will highlight some recent announcements of deployments:**
  - SDSC announced the deployment of Habana technology in their *Voyager* system, incorporating both the training and inference technologies
  - Cerebras Systems deploys their CS-1 platform at the Pittsburgh Supercomputing Center
  - SambaNova has announced deployments at LLNL and ANL, and recently announced an impressive Series D round of investment, making them the most funded AI startup
- **Many of these AI processor companies look to the cloud for their first deployments**
  - Groq, Graphcore, Habana, the two AWS chips (Trainium and Inferentia), and the Google TPU are all available from various CSPs

Conclusions

The future of HPC systems is heterogeneous

- **The growing diversity of HPC workloads is driving diversity of processors and accelerators**
  - With the influx of AI and the incorporation of AI to traditional mod/sim applications, processor and accelerator technologies are becoming more varied in type and focus
- **Some companies are striving for more general accelerator capabilities, while others are focused on specific applications to address with new processor technologies**
- **Future systems are anticipated to be heterogeneous in architecture, with multiple technology platforms all in the same system to address a varied workload set**

HYPERION RESEARCH

QUESTIONS?

anorton@hyperionres.com