Technology Insights... November 9, 2022

J.Gold Associates, LLC. Northborough, MA 01532 USA
www.jgoldassociates.com +1-508-393-5294
Research, Analysis, Strategy, Insight

A Publication For Clients of J.Gold Associates

Intel Goes for the Max

Intel is making a push to recapture momentum in the High Performance Computing and AI space, where it has been lagging lately. Despite having some significant wins in AI processing with its high end Xeon Scalable server platform, it nevertheless has had to fend off competition from AMD EPYC based CPUs, as well as the large market share that Nvidia has captured with its A100, and now H100 AI processors. For its part, Intel still remains the primary choice for high performance CPUs – even capturing Nvidia use of Xeon CPUs in some of its high end system offerings. But Intel's real strategic weakness has been a lack of a competitive high end GPU to combat Nvidia's leadership in that market. That is about to change.

> *"...Intel is making a push to recapture momentum in the High Performance Computing and AI space... We believe that the Max line of Xeon CPUs and GPUs are critical to Intel's long term market success as it pivots to more emphasis on data center and less emphasis on the personal computer markets... Although Intel still has much to prove, we believe the Max line is a strong statement that Intel expects to compete aggressively in this space....."*

Intel Xeon Max

Intel just announced an update to its HPC computing family with the Sapphire Rapids based Xeon Max CPU products, as well as the release of its Data Center GPU Max, based on the Pontevecchio technology it has been perfecting for some time. This is an important announcement for Intel and its user base. A major performance feature is that the products include an optimized High Bandwidth Memory (1TBps) capability, which is addressing a major bottleneck in CPU and accelerator processing.

Claiming a 3.5X improvement over the previous generation Xeon on real world workloads, and as much as 5X improvement over AMD EPYC CPUs, Intel is renewing its leadership in high end processors, albeit at a relatively high power at a 350W TDP. It's a SoC with 56 performance cores, 20 accelerator engines, and the embedded multi-die interconnect bridge (EMIB) connecting the chiplets on the SoC substrate. Despite the high power requirements, Intel claims a 70% reduction in performance per watt compared to an AMD Milan-X cluster. This is not a trivial statistic, as hyperscalers use the total power needed per system licensed as one of the parameters in calculating charges to its customers since power usage is one of the major expenses of running a data center/cloud. Intel claims it already has 30+ systems partners designing Xeon Max products.

As for the new Intel Data Center GPU Max, it claims that it can speed up certain workloads over an Nvidia A100 by 2X. Further, it has 30+ apps that run on Xeon already enabled on the GPU for accelerated processing. The product is available as a double-wide PCIe card (GPU100) with 56 cores, 46GB of HBM and a TDP of 300W, or as modules at either 112 or 128 cores and 96GB or 128GB of HBM at a TDP of 450W or 600W. Finally, it's offered as a 4 module data center subsystem with up to 512 GB of HBM and 1800-2400W TDP, with Intel claiming 15+ system data center design wins. It uses EMIB technology as well as Intel's premiere Foveros chiplet multilayer substrate power and high speed connectivity capabilities. It's expected that products will be available in early 2023.

Real world success

Intel has already established a compelling success. According to Intel, "In 2023, the Aurora supercomputer, currently under construction at Argonne National Laboratory, is expected to become the first supercomputer to exceed 2 exaflops of peak double-precision compute performance. Aurora will also be the first to showcase the power of pairing Max Series GPUs and CPUs in a single system, with more than 10,000 blades, each containing six Max Series GPUs and two Xeon Max CPUs".

"…Cloud and data center continue to grow as more high performance workloads are coming on line. Having high performance accelerated compute is the only way to stay ahead. Indeed, future products from Intel and others will include a combined XPU capability into a single package......"

It's not just about the hardware

One of the most important features of the announcement is not hardware at all. It's that Intel has a "secret weapon" in its oneAPI software. Much of the difficulty in using accelerators is that the application code running the workload may have been written in a proprietary framework/language (e.g., CUDA), which is heavily optimized for a targeted platform/architecture. oneAPI is an open computing model that provides an easy way to build code and then optimize it for a variety of accelerators with minimal intervention. Based on SYCL, it is being deployed by many organizations that need to have cross-component compatibility so as not to get locked into a single vendor's product. This can save weeks or months in workload solutions deployment by not having to rewrite the primary application code each time a new accelerator becomes available. Intel has optimized oneAPI to support its new products, making it relatively easy to optimally port to them.

Is this enough for Intel to recapture leadership?

Given the momentum that AMD has with its data center CPUs capturing increasing market share, and Nvidia has with its GPU accelerators becoming the defacto standard for workloads like AI and HPC, Intel has a lot to prove. It has not been a force in GPUs for many years. But it understands that it must compete effectively in this space if it's to maintain its markets, given that the other companies have leveraged their accelerators to gain ground on Intel in its core CPU product area. Intel does have a strong AI directed effort in place with its Habana Gaudi 2 processors, but this is less general purpose than needed by many workloads like modeling, simulation, graphical/image processing, etc.

The majority of CPU-based workloads at cloud hyperscalers still run on Intel CPUs, despite a growing impact from custom designed and mostly ARM-based processors. And while on-prem data centers remain a strong market for Intel, it has seen erosion in this market as well from agile competitors. Intel must reestablish a presence in the high performance accelerator market, particularly as the HPC and AI markets are growing rapidly. Of course competitors also provide new products to market on a regular basis, so any lead that Intel may have currently will fade with competitive announcements (e.g., Nvidia's H100). It's therefore imperative that Intel commit to continuous improvements in the product line to maintain any leadership it may produce.

Bottom Line: We believe that the Max line of Xeon CPUs and GPUs are critical to Intel's long term market success as it pivots to more emphasis on data center and less emphasis on the personal computer markets that are currently troubled and may remain so for several years. Cloud and data center continue to grow as more high performance workloads are coming on line. Having high performance accelerated compute is the only way to stay ahead. Indeed, future products from Intel and others will include a combined XPU capability into a single package. Although Intel still has much to prove, we believe the Max line is a strong statement that Intel expects to compete aggressively in this space. The next 12-18 months will tell us how successful they'll be in this critical market area.

J.Gold Associates

J.Gold Associates, LLC.
6 Valentine Road
Northborough, MA 01532 USA

Phone:
+1-508-393-5294

Web:
www.jgoldassociates.com

Email:
info@jgoldassociates.com

Research, Analysis, Strategy, Insight

J.Gold Associates provides advisory services, syndicated research, strategic consulting and in-context analysis to help its clients make important technology choices and to enable improved product deployment decisions and go to market strategies.