GIGABYTE – ISC Highlights

At ISC 2022 in Hamburg, GIGABYTE showcased next-generation platforms and liquid-cooled / immersion-cooled solutions for HPC and AI workloads using high-power components from AMD, Ampere, Intel, and NVIDIA. As computing power continues to challenge the Moore's law exponentially in data centers, GIGABYTE is preparing the HPC market with future-proof computing solutions designed with high density of CPU, GPU, DPU, FPGA and more. GIGABYTE offers the largest portfolio of GPU computing solutions in the market and an HCI range of high-density CPU computing solutions, all with optimized air cooling, DLC cooling and immersion cooling in partnership with GIGABYTE's technology partners such as CoolIT, Submer and many others. Find out more with GIGABYTE about how to surpass HPC performance threshold while achieving best energy efficiency.

(Video = https://www.youtube.com/watch?v=T5c9QeFeBLo)

HPC Demo Systems

GIGABYTE demonstrated the platforms: G492-ID0, R282-Z93, G492-PD0, G242-P34, E162-220, E152-ZE0 and more, all of which compatible with NVIDIA GPU, NVIDIA BlueField Series DPU, and NVIDIA SmartNIC solutions. Having physical hardware on demo is always the best for understanding how things work.

The G492-ID0 and G492-PD0 being the showstoppers.

G492-ID0 comes with 8x NVIDIA A100 SXM4 GPU and 2x Intel Xeon Ice Lake CPU and offers the possibility of installing up to 10x NVIDIA SmartNIC to accelerate data transfer

across nodes and clusters and GPUDirect/RDMA. HPC users who work with Artificial Intelligence, molecular simulations, genomics sequencing, weather prediction, and other use cases were impressed by the system performance of G492-ID0.

G492-PD0 and G242-P34, on the other hand, offer alternative HPC solutions to the traditional Intel/AMD-based architecture for programmers and researchers who optimize their software applications using ARM CPU based codes. In particular, G242-P34 comes with a special promotion offer under the “NVIDIA ARM HPC Developer Kit” program for users who are keen to try out the combined performance of NVIDIA A100 and NVIDIA BlueField-2 DPU.

Visitors share their experience of deploying containerized applications, CI/CO, and virtual desktop infrastructure (VDI) using NVIDIA GPU technologies which facilitate virtual GPU performance and resource management by hardware and software features. To this end, GIGABYTE presented the visitors R282-Z93, an all-purpose HPC server model that supports up to 3x GPU (or 5x single-slot GPU) and a broad range of choices for add-on devices such as BlueField Series DPU and ConnectX Series SmartNIC. R282-Z93 provides large system memory and local storage capacities and has a balanced CPU-to-GPU layout (1:1) which is ideal for virtualization.

AIoT and 5G projects were also a hot topic among ISC guests who had technical questions about E162-220 and E152-ZE0. The two server models were developed for the very purpose of lightweight HPC solution deployment at the edge where data collection and analytics must happen in real-time and fast for decision makers. E162-220 is an Intel-based, single-CPU Edge server with support for 1x GPU, whereas E152-ZE0 is an AMD-based, single-CPU Edge server with support for 1x GPU (or 2x single-slot GPU). Both models have multiple expansion slots for installation of DPU, FPGA, SmartNIC and add-on devices which can be required for signal processing and conversion for 5G use cases. The reduced form factor of the two models, coming in as short as 40cm in depth, also fit well with AIoT and 5G project requirements. The commonly referenced GPU models during the event for AIoT and 5G projects were the NVIDIA A100, A30, A10, and A2.

Liquid-Cooled HPC Solutions

GIGABYTE was joined by their partners, CoolIT and Submer, to showcase about how Direct Liquid Cooling (DLC) and Immersion Cooling can have long-lasting impacts on server system performance and CAPEX/OPEX ROI throughout different stages of HPC projects.

As many visitors and solution providers also agreed with GIGABYTE during discussions at the ISC event, they observed a drastic increase in the demand for DLC and Immersion Cooling (mainly 1-phase based) compared to the pre-COVID era. The demand was mainly raised by data center operators and Cloud Services Providers (CSP) who expressed their concerns about incessantly rising computing power and thus the resulting heat output by computing components (especially by CPU and GPU). Conversations also revolved around future EU environmental regulations on energy efficiency of data centers and EU green computing objectives.

GIGABYTE also shared information about next-generation CPU/GPU technologies that will soon bring air-cooling to its limit in terms of thermal dissipation.

Roughly speaking, to use a standard server rack as a basis of comparison, air cooling offers up to 20Kw – 30kW of heat dissipation; liquid cooling offers 80kW of heat dissipation, and

immersion cooling offers 200kW of heat dissipation (with the cooling water temperature below 13°C). The appropriate total solution design varies as every project has specific requirements which require expert consultation. AT GIGABYTE, we help data centers and customers analyze their projects by looking into energy consumption, heat dissipation, space optimization and PUE/ water use efficiency (WUE), among many other technical topics, at each step of solution design.

Taking a step further, GIGABYE also offers installation/deployment services working with data center infrastructure companies, to ensure that customers receive smooth project delivery and short turn-around time for operational readiness. Most importantly, GIGABYTE strongly advises its customers to benefit from its Proof-of-Concept (PoC) resources for validating every solution design and project parameters for making the best decision, as many environmental factors might alter the expected performance and system stability. GIGABYTE has PoC units (in both single-phase and two-phase immersion cooling) for testing and validation of immersion cooled servers. The server model options come in 1U/2U/4U form factors and can be modified on request to suit different use cases and workloads. GIGABYTE works with all major liquid-cooling and immersion-cooling technology partners in the market, so that customers can count on design-in compatibility of GIGABYTE total solutions with their infrastructure.

The conundrum of improving CAPEX/OPEX and FLOPS-per-watt efficiency while surpassing HPC performance targets motivated visitors to spend extra time at our booth for finding the best solution from our HPC solution portfolio.

Development for Next-Generation HPC Solutions

Beyond what was shown by the GIGABYTE team at the event, we revealed tidbits of technical information about what is to be announced in Q3/Q4 2022 for next-generation HPC solutions. GIGABYTE did not focus on product or system specifications but rather on how the new solutions will be able to address end use cases by adapting the system configuration to workflows and data center architectures.

For instance, on the component level, when they look at the newly announced NVIDIA Hopper H100 GPU, they realize that the GPU performance can be brought out only by a great granularity in programming (in terms of thread affinity, applying appropriate computing precisions, cache and memory location and allocation, task distribution, instance isolation, etc.) and in assigning the right resources (hardware wise) to the right type of workload. This sharply contrasts with the usual method of monolithic computing in which the combined computing power matters more so than the computing power for individual tasks which can have different data types and structures, computing models and sequences. Using this example, they explain how to size hardware specifications based on their projects and end use cases. Very often they would propose a mix of high and low GPU density options without regard to where performance bottlenecks might occur through the workflow. Their advice is by no means the golden rule for HPC, but they received positive feedback from ISC visitors who seemed to have found what they were looking for.

Some of the early teasers GIGABYTE gave to visitors about next-generation HPC solutions: the G493 Series (4U 10GPU), the G293 Series (2U 8GPU without PCIe switching), the R283 Series (all-purpose 2U 3GPU), the L-Series (DLC Cooling), and the I-Series (Immersion Cooling). Aside from the speed and performance increase with PCIe Gen5 and DDR5, the highlight was how GIGABYTE pushes the limit of air-cooling effectiveness on high-power components in small form factor server models without capping computing power. Many

visitors were already aware of the new micro-architecture changes in next-generation CPU and GPU, so they were given questions such as how the changes would reflect in performance and costs and how to determine the ROI of investing in next-generation choices. To answer these questions, they looked into the software applications and how codes were structured in the projects of the visitors.

ISC Conclusion

The ISC 2022 brought back excitement about future HPC solutions in the post-COVID era and let GIGABYTE immerse once again in the dynamic of people in the HPC community who continuously seek for better solutions and new ideas of how things can be done. Having their HPC demo systems at the booth generated great interest among visitors in how GIGABYTE's R&D capabilities are realizing GIGABYTE's vision for future HPC solutions that offer both best performance and energy efficiency. GIGABYTE's ever-growing network of technology partners also expedites the development of new solutions and new business models that create new opportunities for users and customers. We are hoping that at the next ISC event we will be able to bring more solutions on live demo and allow visitors to try things out in real time and witness how far GIGABYTE has gone in its HPC ambition.