How Much Can a GPU Withstand?

David Young

Portfolio

Graphics processing units (GPUs) have long been a fundamental aspect of today's computing and display technologies. Their foundation can be traced back over 40 years to the introduction of specialized hardware in early arcade machines, such as video shifters and video address generators, that were capable of drawing basic graphics using scan lines and serial bitmapped video. Advances in manufacturing capabilities, as well as the introduction of the affordable personal computer and 32-bit operating systems, eventually paved the way for the widespread adoption of more powerful add-in graphic processing cards beginning in the mid-nineties. In turn, this preceded the increasing demand for realistic graphics and the centrality of graphical user interfaces that, through technology such as augmented reality and touchscreen displays, is now altering our evolving relationship with machines.

Although the underlying technology behind GPUs has changed dramatically since its early days as simple display controllers, its basic definition has remained relatively the same. Whether existing as a standalone card that plugs into the motherboard, a chipset that is installed directly on the motherboard, or a processor integrated with the CPU, the GPU is designed to manage all graphical functions, such as rendering images, animations, and videos on a display. The methodology behind this design is similar to CPUs in that they both consist of a collection of transistors that manipulate the computer's memory to perform complex mathematical calculations. The main difference, however, is that GPUs process algorithms specifically related to graphics, such as texture mapping and matrix and vector operations.

While GPUs were originally built to realize graphics-intensive purposes, such as gaming consoles and flight simulators (the Whirlwind I, built by MIT in 1951 to train Navy pilots, employed one of the first 3D graphics processors), and have long been associated with high-performance gaming and entertainment displays, significant demand for more advanced graphic display technology is now being seen in the industrial and manufacturing sectors. This development reflects the rapid advances that have been made in displays and human-machine interface (HMI) technology in the past several years. For instance, despite ongoing improvements made to industrial mechanization, cost and implementation concerns meant that the predominant display technology found in factory settings just over a decade ago was still LCDs. However, the intervening years have seen the introduction of more accessible advanced displays, as well as breakthroughs in GPU technology.

The progression of GPU performance capacity illustrates this. Between just the last two generations of

graphics processors, average performance per compute unit grew 325 percent. At the same time, the energy efficiency of each compute unit per watt grew as well, increasing by 262 percent. This has coincided with the emergent widespread usage of smartphones, tablets, and other types of HMI technology, as well as displays capable of high-definition resolutions (a minimum of 480 horizontal lines) and even 4K (upwards of 2,160 horizontal lines). The result of this, in the industrial sector, has been no less than a reconsideration of the way in which humans and machines interact. Whereas LED and other basic displays were once used for simple data collecting and analysis, the current industrial landscape is now using high-definition panels, touchscreens, multi-display workstations, and other setups capable of advanced interactions. In addition, immersive display technologies such as augmented reality are now considered to be in the near future, promising to further disrupt the industrial and manufacturing sectors.

The introduction of more sophisticated graphical user interfaces and displays in industrial settings, as well as the wider availability of graphics processors capable of meeting their demands, is now requiring many industrial organizations to reconsider their current GPUs. However, unlike the consumer market and other non-industrial sectors, these organizations must also contend with additional environmental and logistical challenges when sourcing new parts. For instance, to prevent operating expenditures from becoming excessive, factors such as intelligent power management should be considered to maximize lifetime value and efficiency. Likewise, the stressful conditions of industrial and manufacturing spaces make GPU form factor a vital consideration, while the range of possible industrial applications, both now and in the near future,

require unique scalability of performance. Finally, the increased visibility and vulnerability of many industrial segments necessitate more stringent security standards across the board.

These four factors, when properly considered, can all help decrease total cost of ownership and contribute to the overall efficiency of an industrial organization as they upgrade their manufacturing and display technology, setting them up for ongoing future growth.

I. Intelligent Power Management

Recent innovations in microprocessor designs, including high-speed parallel processing and more compact and integrative architectures on single dies, have led to breakthrough processing speeds capable of the most demanding compute workloads. However, raw processing speeds are not necessarily the ideal measure of a GPU's performance capabilities, especially as it becomes increasingly common to encounter processors combined with heterogeneous hardware blocks, such as CPUs, memory controllers, video encode/decode engines, and more. While the maximum performance capabilities of each component may sound impressive, their aggregate power consumption will likely surpass the maximum threshold allowable for the device, sometimes by several factors. Combine this with the difficulty of writing software capable of efficiently managing the power consumption of each component, as well as the fact that different operations will incur a variety of power requirements, and the need for intelligent power management becomes evident.

This point is an especially important one to make within the context of industrial GPU usage. Although industrial organizations may have more capital with which to source graphics processors, they still must

contend with the ongoing costs of running, maintaining, and replacing parts throughout their lifetime as they are exposed to the conditions of an industrial environment. For instance, variable ambient temperatures, whether high or low, can affect the performance capabilities of a GPU, either making it more likely for the processor to exceed its temperature threshold and overheat, or giving the processor additional thermal headroom that nevertheless remains unutilized. Similarly, the range of various tasks a GPU may be given within an industrial environment, each with different wattage requirements, can make concrete power specifications nearly impossible to determine.

Over time, these deviations can run up operating costs by continuously overclocking processors or, inversely, failing to utilize processing components to their full potential. To avoid these issues, industrial organizations should instead take a more pragmatic approach by utilizing processors capable of bidirectional power management (BDP). This refers to the constant and intelligent monitoring of each component's workload within a processor in order to ensure that maximum performance is reached while still staying inside a specific power management budget. A basic example of this would be a GPU and CPU that have a maximum cumulative power threshold of 30 watts, and so have been allocated 15 watts each. If the GPU is suddenly tasked with an intensive operation, while the CPU remains relatively idle, then BDP will rebalance the wattage to compensate for this discrepancy, taking wattage away from the CPU so that the GPU can perform adequately.

Of course, real-world variables — such as the type of environment in which the graphics processor is functioning, the frequency and intensity of its workload, and the cooling system it uses — may all mean that it is

inefficient to allocate power budgets based on maximum thresholds. For this reason, it is important to source GPUs capable of designing their power budgets based off of realistic worst-case scenarios that take into account the software being used, as well as various external environmental factors. This can all be specified as its thermal design power (TDP).

Essentially the maximum sustained power a processor can use under real-world conditions while still staying beneath defined temperature and voltage limits, TDP is a useful way of increasing efficiency within a known set of constants. However, even this can fail to utilize the maximum available power within a budget if these constants change, resulting in the underutilization of potential GPU performance. To solve this, it is necessary to utilize graphics processors capable of specifying a range of different thermal windows they can operate within, a feature known as configurable thermal design power (cTDP).

By dividing the processor into distinct thermal entities, then using an algorithm that calculates the digital estimate of power consumption for each component and converting these estimates into temperatures, proprietary activity monitors integrated throughout the processor can model current logic activity as an AC capacitance (C_{AC}). This can then be used determine whether any component has reached threshold levels for a given TDP or if there remains available thermal headroom. If a low C_{AC} is causing the processor to consume only a portion of the TDP, then new power budgets can be assigned based on the available thermal window, effectively achieving an ideal power budget. Alternatively, TDP can also be configured manually by system designers based on relevant external factors. For instance, they could specify that they want the GPU to

remain at a high or low TDP for a specific period of time due to variables such as the ambient temperature of the factory floor or additional cooling systems that allow them to disregard normal thermal limits.

For industrial organizations, having this level of control over GPU performance and power management can help make their operations much more versatile and cost effective. The ability to ensure their processors can continue to function within optimal power ranges, regardless of operating or environmental temperatures, is invaluable within the industrial space. Likewise, the option to manually configure TDP enables them to work with a wider range of vendors, including those with products designed for power ranges outside of their TDP, as they can adjust their settings accordingly without sacrificing performance. All this helps produce more efficiency and savings over the long term.

II. Form Factor

Just as the unique conditions of the industrial and manufacturing environment make intelligent power management an essential consideration when sourcing graphics processors, they also affect GPU form factor. Unlike in non-industrial and consumer sectors, in which GPU are often sold as PCI Express cards that are attached to a motherboard, where they are then left undisturbed, GPUs in industrial settings may continuously be exposed to rugged conditions that make such a setup untenable. In the avionics industry, for instance, a processor placed behind a cockpit display must come with the capacity to endure a good deal of shock and vibration without any diminution in performance. Similarly, many other industries will require processors that are small enough to fit into tight spaces, or that come with designs that can be easily

customized according to highly specific ruggedized requirements.

Therefore, industrial organizations concerned with reducing lifetime costs and ensuring their processors can withstand any unique conditions present in their regular operations need to keep the form factor of GPUs top of mind. Specifically, this means looking for GPU designs that allow for maximum configuration within an industrial environment, which, more often than not, will preclude separate graphics cards popular with consumers in favor of discrete GPU chipsets that can be more freely customized. In particular, industrial and manufacturing organizations should look for GPUs built using ball-grid arrays, as well as those packaged as multi-chip modules.

As its name implies, ball-grid arrays (BGA) are a type of surface-mount packaging that consist of a series of solder balls placed in a grid on top of an integrated circuit chip. BGAs replaced pin-grid arrays, which were previously used to mount microprocessors, since the solder balls can be more easily placed in a denser array that allows for greater interconnectivity. The benefits of this design within an industrial context are numerous. In addition to the increased performance capabilities that BGAs make possible, they are also more efficient at thermal conduction, have a lower inductance due to the shorter distance between the processor and the circuit board, and help reduce the overall thickness of the GPU package, allowing for versatility and configurability. All of this is important for industrial organizations in need of high-performance solutions capable of withstanding high ambient temperatures, increased wattage needs, and rugged conditions.

The small-footprint and high-endurance package that is possible when designing GPUs using BGAs makes them a perfect solution for multi-chip modules (MCM). Broadly defined, a MCM refers to an electronic package

that consists of multiple integrated circuits, as well as other components, that act in concert on a single device. They can be built by combining multiple processors on a single substrate, by connecting them together using wire bonding, or by numerous other means. Regardless of their build-type, one of the greatest advantages of this package is that, while individual chips offer limited customization options, MCMs can be built according to custom specifications. For instance, after accounting for an organization's performance requirements, the MCM package can then be custom designed to fit within specific architectural limitations, to integrate with protective equipment (such as vibration dampeners), and more.

MCMs also allow customers to streamline the process of sourcing multiple components. This makes it much easier for them to simplify their supply chain, for example, by utilizing a single supplier for both their GPU and memory needs. However, it should be noted that none of this comes at the expense of versatility. A customer can still source whatever components they deem best suited for their requirements, while also ensuring they can be compactly packaged and designed according to their own specifications, rather than those of a manufacturer. This makes them an ideal solution for industrial organizations in need of reliable and rugged GPUs.

III. Performance Scalability

The raw performance capabilities of modern-day graphics processors is increasingly becoming a point of concern for organizations within the industrial and manufacturing sectors, especially as more of them adopt advanced display technology like high-definition screens and multi-display configurations, and move into

next-generation systems such as machine vision and augmented reality. However, as this is still an ongoing, iterative process for many organizations, it is just as important to take into consideration the scalability of current GPUs. Although large segments of the industrial sector may be posed to take advantage of the latest display and HMI technology, their current needs may still be limited to only basic functions. What's more, as they begin the process of scaling up their GPU capabilities, it should not be necessary for them to replace their entire software infrastructure in order to accommodate additional functionality. What all this means is that, for industrial organizations sourcing graphics processors for both current and future use, the importance of adopting GPUs that embrace open standards cannot be understated, while the ongoing convenience of integrative solutions should be noted as well.

Open standards, when being used with regard to GPU hardware, refers to open-source kernel languages such as OpenCL, open-source graphics APIs such as OpenGL, and other royalty-free open standards that enable the ongoing development of cross-platform technology. These open-source frameworks make it possible to build and operate programs that can execute across a heterogeneous array of GPUs, CPUs, hardware accelerators, and other hardware components and devices, independent of whatever resources are available. In other words, they let industrial organizations and their developers adopt graphics-processing solutions capable of seamlessly integrating into their existing hardware and software infrastructure, allowing them to upgrade their graphics capabilities with as little disruption as possible.

An additional advantage that open-standard GPUs can give developers is the opportunity to leverage their

capabilities across not only as wide a range of applications as possible, but also across applications not immediately connected to graphical uses. A pertinent example of this would be machine vision. A highly integrative field that depends on a cooperative ecosystem of different technologies — including cameras, image-processing engines, and GPUs capable of robust data management — machine vision demonstrates how graphical processing, when combined with open standards, can effectively take graphics and merge intelligent data processing with it.

Apart from open standards, another way industrial organizations can ensure they are adopting GPU solutions that will deliver the performance they demand while also giving them the ability to effortlessly scale up as their needs change is with integrated solutions, which are also sometimes referred to as accelerated processing units (APUs). These solutions combine a CPU, GPU, and I/O on a single chip, reducing the overall form factor and simplifying the design by eliminating the need to acquire additional processing components to the system. This makes integrated solutions like APUs distinct from MCMs, as the former are designed to deliver a streamlined balance of power and performance within a specific environment, while the latter are a custom-built combination of GPUs with a memory subsystem that still must be paired with an external controller.

The convenience of APU-like solutions is evident by how much the tight integration they offer between the GPU and CPU helps maximize the amount of performance one can get from either. For example, the combination of a GPU core on the same die as a CPU makes it possible to more efficiently distribute intensive data-processing workloads from the CPU to all available processor cores in parallel, increasing image-processing performance by, in some cases, several orders of

magnitude. The integrated design also simplifies other architectural elements, such as power voltages and power rails, further increasing overall performance while making it possible to maintain much smaller form factors, opening up the possibility of even more impressive versatility and customizability. And, while it may be necessary to add memory onto the system as upgrades are made, this is still much easier than adding in additional CPUs or GPUs, as the case may be.

Industrial organizations of every type will be able to draw benefits from maintaining open standards in their GPU hardware, especially as they seek to upgrade and integrate their existing hardware and software infrastructure with the demands of more intensive graphics processing. Integrative solutions like APUs, meanwhile, will remain an attractive option for industrial organizations with more established applications and environments, such as those in avionics, in need of simple, reliable, and scalable graphics processors.

IV. Security

An unavoidable concern in today's increasingly connected world, security is an especially important issue to address for industrial and manufacturing organizations, particularly those that deal with sensitive information (such as the medical industrial organizations) and those with high visibility (such as prominent manufacturers). Hackers and viruses have reached an unparalleled level of sophistication in recent years, as is evident by the raft of high-profile ransomware and other illegal digital infiltrations that have made headlines across Europe and North America. For this reason, even when sourcing new GPU technology, it is vital that organizations check that it comes with robust security features that reduce the

chances of unauthorized programs and individuals from gaining access through vulnerabilities. To do this, industrial organizations should look for GPUs that come with hardware-based encryption techniques.

Every time a device boots up, it goes through a process in which it must validate its own boot code before it can begin properly functioning, a task that it can either complete through software-based or hardware-based encryption. Although software-based encryption can be useful for applications that require selective encryption of certain files or directories, it suffers from numerous disadvantages when compared to hardware-based encryption. Consider the methods by which both encryption techniques function: Whereas software-based encryption runs on the operating system, hardware-based encryption bypasses all other devices and systems and is able to validate the GPU boot code using only the hardware itself. The device wakes up, retrieves its keys, fetches the code, caches the operating-system images, compares caching to what it should match with, then, if it interprets the operating image as valid, releases the x86 cores.

The benefits of this approach give hardware-based encryption a much higher level of security. For instance, because the physical hardware is doing the validation, it allows its applications to boot into a trusted environment surrounded by a firewall. This also prevents the possibility of someone inserting malicious code into the system and interfering with the boot process. Furthermore, by conducting all encryption and validation on the hardware accelerator itself, much less power is consumed as compared to software-based encryption, which must be run on the CPU core. Hardware-based encryption can even be configured to securely execute crypto-coprocessing across an ethernet, enabling encryption from one device to another.

While no organization may be able to completely protect themselves against emergent digital threats, hardware-based encryption offers one of the most complete security strategies available. Especially for those industrial organizations that regularly deal in sensitive data or that are prominent targets in and of themselves, the act of upgrading to a more modern graphics processor can also be an ideal opportunity to enhance their security.

The increasing importance of GPU technology in the industrial and manufacturing industries should come as no surprise. Graphics processors have been advancing at a remarkable pace for years, growing from simplistic video display controllers in their first iterations to the variety of powerful graphics boards, discrete chipsets, and integrated solutions available today. Along this path, they have helped lay the foundation for continuous parallel improvements in display technology, popularizing high-definition displays and touchscreens, and transforming the way humans interact with machines. It was only a matter of time before these developments made their way to the industrial and manufacturing sectors and began changing them as well.

Where there was once demand for little more than basic LED displays, industrial organizations are now in the midst of upgrading their facilities with workstations capable of a range of graphics-intensive applications, with some even looking ahead to the possibilities inherent in near-future technologies like machine vision, augmented reality, and artificial intelligence (AI) capabilities. With this changing landscape also comes an increased demand for more capable graphics processors that can meet these new requirements while accommodating the unique needs of the industrial space. It is an exciting time for new developments, but also a

good time to take stock and examine the most important considerations when sourcing new GPU technologies.

Moving forward, the ability for modern-day graphics processors to intelligently manage their power requirements by monitoring their individual elements and adjusting their voltage requirements based off of configurable thermal windows is essential not just to processing efficiency, but also to long-term costs. The same argument can be made for the overall form factor of current and next-generation industrial GPUs, all of which must be able to withstand environmental stresses while remaining versatile enough to be customized according to whatever a particular industry demands. Scalability, which includes a processor's ability to function seamlessly within an existing organization's infrastructure, will also remain an important concern as long as GPU technology continues to advance. And, of course, the security of this technology, as well as the integrity of the data it protects, can never be overlooked.

All of these are essential for today's industrial and manufacturing organizations to keep in mind when attempting to select the most effective GPU for them. Doing so will help successfully deliver the performance requirements they desire, while also decreasing their total cost of ownership and setting them up for sustained growth.