IDC
Analyze the Future

I D C A N A L Y S T C O N N E C T I O N

Peter Rutten
Research Manager

The Value of Purpose-Built AI Workstations

January 2018

Applications based on artificial intelligence (AI), machine learning, and cognitive computing, which use unstructured data and information as their fuel, require compute infrastructure with capabilities that range from above-average performance to accelerated performance. Therefore, IDC believes there is a case to be made for purpose-built AI workstations.

The following questions were posed by NVIDIA to Peter Rutten, research manager, Server Solutions with IDC's Servers and Compute Platforms Group, on behalf of NVIDIA's customers.

Q. Why does an organization need an "AI workstation?"

A. Businesses have decided that AI is critically important for their competitiveness. They don't always know yet what kind of AI, and they often don't know how to get started, but they're freeing up budget, and they're tasking developers to start experimenting. Larger businesses can afford to recruit skilled developers, and smaller businesses ask their current team to develop the necessary skills. We're seeing a lot of trial and error among these businesses right now, and there seems to be room for that.

Nevertheless, there's also a distinct urgency that drives these companies: They are anxious to not fall behind. Developers are working hard to get a good grasp of the deep learning (DL) frameworks, growing their data sets, and starting to develop models. What I've noticed is that it is quite beneficial if developers, at this stage in their AI journey, don't need to rely on IT and can be freed from having to obtain access to a shared datacenter server, not in the least because GPU-accelerated servers aren't a dime a dozen in the typical datacenter. Developers need free rein to bring this important new AI capability into the business, and the best way to achieve this is by giving them their own tools.

Of course, those tools need to be suitable and performant enough for such heavy-duty tasks as DL. A workstation with multiple powerful GPUs and a well-integrated, fully optimized stack is ideal for developers in this experimental stage. It allows them to get to production-scale deep learning much faster, which means that the business will start benefiting sooner from the models they develop.

Q. Doesn't a purpose-built AI workstation seem like an expensive proposition compared with simply installing four high-end GPUs in a workstation enclosure?

A. In terms of capital expenses, adding a bunch of GPUs to a workstation is less costly than purchasing a fully integrated AI workstation solution. But the never-ending theme with regard to procuring compute is: What are the operational expenses you incur when you try to lower your capital expenses? I'm seeing this again and again — the quick decision to go for the cheapest solution in terms of initial outlay, which then comes back to bite you.

In this case, the developers who are tasked with assembling a solution will be spending an inordinate amount of time on trying to make that configuration work rather than building the AI models that the business is waiting for. They'll be specifying and assembling pieces of hardware and software and then troubleshooting any incompatibilities they run into. They'll be modifying code in the open source DL framework they've selected to make it work properly with the hardware. And once things are up and running, they'll be searching for answers from vendors or community forums whenever there's a software update or when a hardware component fails.

All in all, we're probably talking weeks, if not months, of their valuable time. Some hardware vendors assert that hardware and software are so completely integrated in their workstations that developers will not be wasting their time on any of these activities. These vendors may have invested an enormous amount of expertise around stack development, integration, and optimization on GPUs and have developed their product with the following scenario in mind: The workstation arrives and is placed under a developer's desk; the developer plugs in the workstation and within hours is doing actual AI model work.

Q. Isn't a GPU server in the datacenter more efficient as a deployment methodology?

A. Initially, probably not for all the reasons touched on previously. Putting an 8-GPU server in the datacenter has scale and efficiency advantages, but it may well be overkill at this stage for many organizations that are just starting to develop AI models. They may simply not be utilizing all those GPUs yet, and that would be a waste of money. There are server solutions with fewer GPUs, and some even come with a software stack for DL, but that takes us back to the previously discussed disadvantage of forcing the developer to share the resource with other users across the organization and to negotiate with IT regarding its configuration.

Again, there's something to be said for giving these developers free rein. A 4-GPU workstation is a nice midway point — half the GPUs at half the cost of a comparable 8-GPU server. And the developers can always decide to share the workstation among themselves by remotely accessing it. That's a good model for smaller businesses — they can justify the cost, ensure maximum utilization, and build their AI models on the workstation until they're ready for prime time in the datacenter.

Q. Acquiring personnel with AI expertise from a small global talent pool is challenging for most organizations. What can organizations that are just starting to build their practice do to attract these highly valued individuals?

A. True, the pool of AI talent is small. As a result, AI experts command top salaries, and they can be picky about where and in what kind of environment they want to work. They're also very passionate about their work, and they get excited about tools that help them do their work better, faster, more flexibly. That's why many of them want to work for the likes of Google or for well-funded academic organizations that are building supercomputers. Having access to the best tools is an important motivator for these people.

So how do you compete for this talent as a smaller organization? Mostly by understanding what workstyle and tools AI experts feel will empower them. They don't want to be slowed down, waiting for a model to run its course, when they know that there are tools that would allow them to do it much faster. They don't want to have to negotiate with IT. They don't want to have to build their own environment. They want to get to work right away on the stuff that really matters. A multiple GPU–accelerated workstation with a fully integrated and optimized AI software stack under their desk will give them the compute power they desire. That's a great start to get them interested in working for you.

ABOUT THIS ANALYST

Peter Rutten is a research manager, Server Solutions with IDC's Servers and Compute Platforms Group. In this role, Mr. Rutten focuses on the high-end server market and the accelerated server market. His research on the high-end server market includes RISC-based systems, mainframes, and mission-critical x86 platforms and their operating environments (Linux, Unix, z/OS). His research on the accelerated server market includes servers with GPUs, FPGAs, ASICs, many-core processors, and other accelerators that are used for AI or for highly demanding workloads. Furthermore, Mr. Rutten covers infrastructure for select workloads such as AI and SAP HANA.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com