Sponsored by: IBM

Authors:
Peter Rutten
David Schubmehl

September 2017

Hitting the Wall with Server Infrastructure for Artificial Intelligence

IDC OPINION

Businesses are struggling with numerous variables to determine what their stance should be regarding artificial intelligence (AI) applications that deliver new insights using deep learning. The business opportunities are exceptionally promising. Not acting could potentially be a business disaster as competitors gain a wealth of previously unavailable data to grow their customer base. Most organizations are aware of the challenge, and their lines of business (LOBs), IT staff, data scientists, and developers are working to define an AI strategy.

IDC believes that this emerging environment is to date still highly undefined, even as businesses must make critical decisions. Should businesses develop in-house or use VARs, systems integrators, or consultants? Should they deploy on-premise, in the cloud, or in some hybrid form? Can they use existing infrastructure, or do AI applications and deep learning require new servers with new capabilities? We believe that many of these questions can be answered by starting with a well-coordinated small initiative on-premise and then scaling it while keeping a close watch on the impacts.

At some point, businesses taking this trajectory will experience what businesses more advanced with AI applications have already been exposed to: they will hit the wall with their server performance. AI applications, and especially deep learning systems, which parse exponentially greater amounts of data, are extremely demanding and require powerful parallel processing capabilities, and — increasingly — it seems evident that standard CPUs cannot sufficiently execute these AI tasks. Early-stage and advanced-stage AI users will at some point have to overhaul their infrastructure to achieve the required performance capabilities.

IDC therefore recommends that businesses that are developing AI capabilities, or scaling existing AI capabilities, should "hit the wall" in a tightly controlled fashion. Do it knowingly and in full possession of the details to make the next infrastructure move. Also, we recommend they do it in close collaboration with a server vendor that can guide them from early stage to advanced production to full exploitation of AI capabilities throughout the business.

The sections that follow in this white paper were informed by an extensive IDC survey among 100 adopters of accelerated compute infrastructure for AI applications in North America as well as by findings from 8 in-depth interviews with organizations that are running AI on accelerated compute.

SITUATION OVERVIEW

The Variables of Embracing Artificial Intelligence

Businesses around the world are responding vigorously to the new opportunities offered by AI workloads. AI workloads include applications based on machine learning and deep learning, using unstructured data and information as the fuel to drive these applications. Some businesses are well on their way with deploying AI workloads, others are experimenting, and a third group is still evaluating what AI applications can mean for their organization. For all three stages of progress, the variables that, if addressed properly, together make up a well-working and business-advancing solution are numerous.

To get a handle on these variables, executives from IT and LOBs in many businesses, sometimes in the form of special committees, are actively considering their organization's approach to the AI opportunity. One fundamental question is: What is the business purpose of the AI initiatives we are contemplating? This is an important question (no one wants to invest in AI for the sake of AI), but there is no need to reinvent the wheel. Many well-defined use cases exist that are applicable across industries. For example, IDC has identified a set of AI use cases that include

- Fraud analysis and investigation (banking and other industries)
- Program advisors and recommendation systems (many industries)
- Regulatory intelligence (many industries)
- Automated threat intelligence and prevention systems (many industries)
- IT automation (most industries)
- Sales process recommendation and automation (retail and other industries)

- Diagnosis and treatment (healthcare)
- Quality management investigation and recommendation (manufacturing)
- Supply and logistics (manufacturing)
- Asset/fleet management (most industries)
- Freight management (transportation)
- Expert shopping advisors and product recommendations (retail)

Many of these use cases can be developed in-house or are available as commercial software. Some are also available as SaaS in the cloud. This leads to the second major variable — when an organization has identified a suitable AI use case, the next question typically is: Do we develop in-house, get an off-the-shelf solution, engage with a third-party packager, or identify a cloud solution? Developing in-house can be challenging, but is not unsurmountable, and potentially rewarding. IDC has found that 23% of businesses don't know what the right software or algorithms would be for the AI solution they're considering. This may seem like a significant percentage but it also shows that a majority of businesses have been able to identify the right algorithms.

The advantage of developing an AI application in-house is that the solution will be finely tailored to the business needs, as opposed to, for example, a cloud solution. Also, the approach avoids the cost of using third-party services: 32% of businesses have found that outside services are too expensive. What is less clear to many businesses is whether they should use open source frameworks or commercial software for developing their AI solution. Cost plays a role here too because 31% of businesses find the cost of cognitive software from industry leaders too high; instead, they choose from the rich variety of open source frameworks that are available as downloads or prepackaged with the server they've procured for their AI initiative.

Often, the next topic of consideration is whether the business has the right data for an AI solution to be effective. Here too, IDC has found that most businesses seem to have a good understanding of the data required, but a fourth of businesses struggle with preparing the data, including data cleansing, labeling, and transforming, which is very resource and labor intensive, as well as with managing the sheer volume of data that is fed into the AI application. Keeping sensitive data that is being used for training the AI solution secure is also a concern. Support with data preparation can be had from various providers, while choosing the right server hardware plays a decisive role in the ability to manage data volumes securely.

And, of course, the discussion turns to the skill sets that are required and whether they are available. In many cases, developers train themselves to become versatile with the AI frameworks, and this has proven to be a successful approach. Hiring engineers with fully developed AI skill sets can be expensive: Almost a third of businesses say that the cost of staffing for AI software is too high. The same is true for data scientists, who are needed for more complex solutions; they can be hard to find and are typically expensive to recruit. It is therefore not unusual for businesses to get started with an AI initiative via an evangelist in the company who energizes developers to obtain the needed skills, sometimes simply through experimentation, and who engages with the infrastructure team to carve out a suitable part of the environment to develop, test, and deploy new AI applications.

On top of these variables to tackle, there is the question of what server infrastructure businesses need for their AI initiatives, which will be the focus of this white paper from here on. Questions that arise with regard to infrastructure are whether an AI initiative can be run on existing infrastructure (we believe that — briefly — that's an acceptable start) or whether new infrastructure is needed (ultimately, yes) and what that new infrastructure ought to be. And obviously, businesses are also asking themselves whether they can run their AI solution in the cloud.

The Diversity of AI Applications in Use Today

Many businesses are eager to bring AI capabilities to the organization. IDC research shows that by 2021, vendor revenue from cognitive software and cognitive server infrastructure will grow to $10 billion and $9 billion, respectively, and businesses are betting on this fast-emerging technology to develop new, competitive capabilities. A few examples from IDC's in-depth interviews with organizations that have AI applications in production or testing illustrate the diversity of solutions and deployment scenarios:

- A midsize real estate company uses CognitiveScale for analyzing property sensors data. It also uses Dato (formerly Turi), which is a recommendation engine for fluctuating property rents, as well as PTC ThingWorx and GE Predix and industrial PaaS that includes machine learning models for detecting anomalies, directing controls, and predicting maintenance.
- A large bank uses AuthenticID for face detection, speech recognition, and sentiment analysis, as well as IBM Watson for cybersecurity and for determining customer behaviors and patterns and creating personalized offerings. The bank is also conducting a proof of concept (POC) with Marstone, which is a robo-advisor and financial planning platform, and with Saffron, a solution for enabling small devices to perform intelligent local analytics.

- A midsize healthcare company has developed various apps in-house: Alerter, which is a machine learning app that identifies issues with machines, and Responder, which automates the remediation of these issues without human interaction. The healthcare company is also developing a deep threat intelligence learning application to be offered as a service and is using HPE Haven as well as cognitive services from Microsoft.
- A hospital in Thailand uses IBM Watson for Oncology to provide doctors and medical staff with comprehensive information about cancer cases by analyzing each patient's data against thousands of cases, including information from 5,000 hours of training by oncologists at Memorial Sloan Kettering Cancer Center, New York; 300 medical journals; 200 textbooks; and 12 million pages of text.

Hitting the Wall with AI Infrastructure

IDC has found, however, that most businesses that are in POC or production mode with AI and deep learning applications have at some point hit what we refer to as "the infrastructure wall" — sometimes not once but twice after they had moved to different infrastructure. Hitting the wall is a term from endurance sports, where an athlete is confronted with sudden fatigue and an immediate and dramatic loss of energy due to glycogen depletion. As a metaphor of what businesses are experiencing with their infrastructure for AI workloads, it is quite apt.

IDC asked organizations what they experienced when they started running AI applications on their existing on-premise infrastructure, and the responses were stark. 77.1% of respondents said they ran into one or more limitations with their on-premise AI infrastructure. Among cloud users for cognitive, a remarkable 90.3% of organizations ran into such limitations. Table 1 lists on-premise and cloud infrastructure limitations with AI apps.

TABLE 1 Infrastructure Limitations with AI Apps (Ranked in Order of Prevalence)

On-Premise Infrastructure Limitations with AI Apps	Cloud Infrastructure Limitations with AI Apps
Difficult to manage	Difficult to scale
Difficult to scale	Performance limits
Performance limits	Insufficient storage
Cannot complete tasks within the SLA	Difficult to manage
Insufficient storage	Difficult to diagnose problems
Difficult to diagnose problems	Difficult to balance load
Server virtualization difficulties	Cannot complete tasks within the SLA
Lack of interoperability in the datacenter	High energy use
High energy use	Lack of interoperability in the datacenter
Memory limitations	Server virtualization difficulties

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

Because of these difficulties with their infrastructure, businesses go through generational shifts quickly. AI applications and deep learning have only been around for a few years, but IDC has found that already 22.8% of businesses are on third-generation infrastructure for AI applications, while 37.6% are running on second-generation infrastructure and 39.6% are on first-generation servers. These percentages are indicative of a search for the right infrastructure. Table 2 lists the most often occurring generational shifts for AI server infrastructure.

TABLE 2 Most Common Generational Shifts for AI Server Infrastructure (Ranked in Order of Prevalence)

Moving to greater processor performance
Moving from scale out to scale up
Moving from a VM to a dedicated server
Moving from scale up to scale out
Bringing in greater I/O bandwidth
Moving from a dedicated server to a VM
Adding accelerators

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

Moving to a system with greater processor performance (the most common action taken), greater I/O bandwidth, and accelerators is a logical decision. But this data is also showing that there is uncertainty about the ideal configuration. Some businesses have tried scale-out and moved to scale-up; some businesses have done the reverse. Other businesses started in a VM and then moved to a dedicated server, while some of their peers did the opposite.

These contradictory moves are not as strange as they may seem. Businesses are experimenting not just with the AI software but also with the infrastructure to run it on. Some businesses started on a scale-out configuration and, as their solution matured, decided that they needed more performance, which they found on an existing scale-up system in their datacenter. Other businesses started a POC on a partition of a scale-up system, and upon taking the solution to the next stage, they decided to move it to a cluster of one- or two-socket servers. Similarly, a solution may have been developed in a VM and then migrated to a dedicated server to be developed further in a somewhat insulated environment (something many businesses like to do in the early stages).

IDC believes that for early experimentation and development, all these moves make sense. Leveraging the existing environment means delaying investing in new server infrastructure until it has become clear what the right configuration should be. However, once an application is getting close to up and running and being readied for production, sound infrastructure decisions need to be made to avoid hitting the infrastructure wall.

Based on responses from businesses that have been running AI applications, we believe that the ideal infrastructure configuration for cognitive applications is a cluster of one- or two-socket servers with accelerators, although accelerators may also be added at a later stage as the need arises. A cluster of midsize systems are also viable, but they would only be relevant in the case of a workload that is scaling very rapidly. Other configurations may be feasible. What appears clear from research among users is that hyperconverged systems and VMs have proven to be less effective for cognitive applications.

What Should You Do?

IDC believes that businesses that are currently considering AI initiatives or that are moving from an experimentation stage to a more mature stage can take any or — over time — several of the AI development approaches discussed in the sections that follow.

Small to Medium-Sized AI Initiatives

For small to medium-sized AI initiatives, developing a solution in-house is recommended. There are multiple advantages with this approach. Through collaborative experimentation, developers, LOBs, data analysts or data scientists (if available), and the infrastructure team will obtain important new skill sets while creating a tailored solution for the business. Data analysts and data scientists can prepare data sets and the related models, developers can test frameworks, the infrastructure team can evaluate on what hardware to develop and what to use for production, and the LOBs will have an opportunity to set the parameters that the solution should fulfill. However, it is advisable to take this approach only for unique AI projects. If the desired solution is readily available as commercial software, the benefits of in-house development will be outweighed by the business benefit of rapid deployment that a commercial package allows.

IDC recommends starting small and on-premise. The tendency will be to start on a dedicated server somewhat insulated from the rest of the environment, but be aware that integration will ultimately be important. If there is an AI training component, then the environment will need to be able to access the data that is intended for the training and the hardware will need to be capable of strong parallel processing, ideally with a sufficient number of accelerators, such as graphics processing units (GPUs). The environment can consist of a cluster, which AI solutions tend to prefer, and even a converged system with multiple nodes. However, for first-generation AI infrastructure, a hard partition in a scale-up server can work as well. VMs or hyperconverged systems are less suitable. If the data is business critical, then a hard partition in a scale-up enterprise-class server that hosts the data might be useful as the organization will not need to move the data out of its secure environment. Note that there is a wealth of open source frameworks for AI development that only runs on Linux.

Once the infrastructure team, development team, and data scientists are comfortable with the solution, have run the solution in production, and have experienced the capabilities and limitations of the software and the hardware, the business will be much better able to determine the next steps. Those next steps may include continuing building out on-premise, in-house capabilities, including upgrading or expanding the infrastructure, adding a cloud component, and/or bringing in others such as VARs or consultants.

It is critical that during this trial-and-error stage, the infrastructure team thoroughly investigate new infrastructure solutions. As previously mentioned, AI systems run well on clusters of single- and dual-socket servers with high per-core performance and I/O parameters combined with accelerators such as GPUs. The team should not only consider server products available from its traditional vendor but look at other server vendors as well, especially those that are offering a complete AI hardware/software stack. Some of these vendors provide help at all stages of the deployment of an AI system, from hardware selection and optimization through the software stack all the way to deployment and consulting services. It is recommended that a vendor that has demonstrated deep knowledge of infrastructure requirements for AI and deep learning is selected.

Make sure that the vendor can advise on the first experimental stages, even if that is on existing hardware, and can then guide the organization toward on-premise or a hybrid on-premise–cloud expansion. Ideally, the vendor can work through several or even all the small to large scenarios; in other words, serve as an advisor for the small initiative but also as a consultant to the next stage — a larger AI initiative.

Larger AI Initiatives

Larger AI initiatives will benefit from external support. The time, cost, and complexity of developing a comprehensive AI solution that is intended to bring business-critical innovations to the organization may be too great to take on with an in-house trial-and-error approach, except for large organizations with significant resources. Third-party AI solution providers can help implement a solution rapidly, as can VARs or systems integrators, but they will be less flexible and less tailored to unique business needs. Very large initiatives can benefit from a consulting partner. Consulting partners tend to be expensive and create long-term dependencies, and the initial deployment time is typically long. On the other hand, the resulting solution will be fully tailored to the organization's needs and, if executed properly, well integrated with the datacenter.

For large initiatives, working with a server vendor with AI expertise and a range of AI offerings that encompass the entire hardware/software stack also has distinct advantages. The server vendor will generally be less expensive than a third-party consulting partner and more knowledgeable about optimization and scaling of its own hardware than other solutions providers. The latter is not a trivial point — make sure that the vendor has a demonstrated ability to scale infrastructure for AI applications and deep learning because scaling accelerated compute nodes is not as straightforward as scaling compute nodes with just CPUs.

Source: IDC, 2017

It is recommended for the LOBs, the development team, and the infrastructure team to remain intimately involved to make sure the AI solution is customized as much as possible and that skill sets will be developed (through training). It is important to ensure that the business does not end up with a "black box" solution that only the server vendor or solution provider understands, that doesn't scale well, that doesn't integrate with the datacenter, and that runs into performance limitations when transaction or data volumes start to increase. In other words, none of these approaches will make the infrastructure team's task any easier. AI server vendors, solutions providers, and consultants will be making hardware recommendations that should be critically reviewed with regard to the same parameters as with in-house development: accelerated performance, I/O, manageability, and scalability.

Note that several of these scenarios can be combined, in terms of both approach and deployment. For example, a solution built in-house can be combined with a SaaS solution in the cloud to achieve a hybrid solution, or a solution built in-house can be followed by a larger implementation by a VAR. Last, IDC has found that most organizations do not have clear estimates of cost for infrastructure or software for their AI efforts. Businesses need to develop metrics for AI projects, including the cost of software, infrastructure, and labor. They should also calculate the potential for payback (through either improved productivity, reduced costs, or increased revenue) and make sure that they collect data on these metrics as the project gets under way.

On-Premise or Cloud?

For some larger AI initiatives, SaaS solutions may exist, but as with any cloud-based software solution, customizability will be limited and scalability will depend on the provider's infrastructure, as will performance. Also, cost can become detrimental when data volumes or the number of transactions grow rapidly. In the case of business-critical data, data that is sensitive, or data that is subject to regulatory compliance, the security of a SaaS solution will need to be evaluated.

IDC has found that among businesses with accelerated infrastructure for AI applications, 65% run these solutions on-premise: 22% on-premise only and 43% both on-premise and in the cloud. A majority of businesses say they have found the cloud experience to be satisfactory so far and will move AI workloads to the cloud. Yet this migration will not affect the overall distribution of cognitive workloads across all possible deployments in the next 24 months; in other words, the percentage of on-premise deployments remains the same. Nor are certain AI use cases deemed more suitable for either on-premise or cloud, with some exceptions. AI use cases such as diagnosis and treatment, for example, tend to be more prevalent on-premise than in the cloud, thanks to data security concerns. Merchandizing for omni-channel operations, however, has somewhat higher prevalence in the cloud. Nevertheless, there is a clear role for on-premise, cloud and, of course, a hybrid strategy. The latter is likely to become the most advantageous deployment approach.

Accelerators

In this white paper, we have on various occasions mentioned accelerators as an important way to overcome infrastructure performance limitations with AI systems. This section therefore briefly discusses accelerators. This is especially true with AI systems that employ deep learning algorithms, which require massive compute capabilities to train. In some cases, training deep learning algorithms with accelerators can bring iterations down from days to hours.

Per IDC's definition, accelerated computing is the ability to accelerate applications and workloads by offloading a portion of the processing onto adjacent silicon subsystems such as graphics processing units and field-programmable gate arrays (FPGAs). Accelerated computing is gaining traction in the enterprise as businesses seek solutions for overcoming the limitations of CPUs for workloads, such as AI applications.

GPUs are especially attractive to businesses as they can be procured off the shelf and utilize standard libraries that can easily be incorporated into applications. However, other technologies that offer potentially higher performance per watt such as FPGAs, many-core processors, and application-specific integrated circuits (ASICs) are starting to gain traction as well:

- A GPU performs vector and matrix computations that underlie neural network layers. GPUs do so in a parallel way, providing vastly improved training speeds with better energy efficiency.
- A many-core microprocessor is optimized for parallelism and/or vectorization without the use of an external accelerator. A many-core microprocessor has more cores than a typical multicore CPU and is part of an architecture that aims to maximize data transfer rates between the processor, cache, and memory. It also performs the traditional functions of a CPU.
- A coprocessor is a PCIe card used to accelerate parallel workloads. It incorporates a many-core processor and includes dedicated cache, memory, and an operating system kernel but needs a CPU to bootstrap.
- An FPGA is an integrated circuit designed to be configured by the customer after manufacture, using a hardware description or high-level language. FPGAs are composed of an array of programmable logic blocks, interconnects, and I/O blocks. They can also be reconfigured.
- An ASIC is a purpose-built integrated circuit that cannot be reconfigured after manufacture.
- An interconnect is a data connection between a GPU, an FPGA, or an ASIC and a CPU. A PCIe interconnect has a maximum one-way bandwidth of approximately 16GBps, while NVIDIA's NVLink 2.0 has a maximum one-way bandwidth of 150GBps.

Source: IDC, 2017

Most smaller companies elect to buy accelerators as part of a server from a server vendor. This is a convenient approach as most prominent server vendors have an accelerated server offering. Larger companies also go to VARs or systems integrators or buy directly from the accelerator vendor. This approach provides them with more flexibility as VARs and systems integrators will be able to deliver a more customized solution, while buying directly from the vendor provides complete flexibility to install accelerators.

When buying accelerators as part of a server, expect a price premium. There are, to date, few benchmarks to determine how much extra performance an accelerator delivers as part of a given server, but IDC research shows that businesses that have procured such systems on average find the price premiums acceptable for a given total performance increase (see Table 3).

Acceleration is very effective but not always the ultimate solution to infrastructure limitations. Much depends on the core performance of the server, the type of acceleration chosen, the type of interconnect, and various other factors such as software and data. It is therefore imperative that businesses not just consider which accelerators and how many but also in what kind of servers they are to be installed, including their per-core performance and I/O bandwidth. It is key to select a balanced system, especially for organizations that are in the experimentation stage with AI trying various models, as each model will stress the system differently.

TABLE 3 Acceptable Price Premiums for Given Performance Increases

Performance Increase (%)	Price Increase (%)
25	19
50	25
75	31
100	36

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

FUTURE OUTLOOK

IDC expects AI applications to not just proliferate fast as distinct solutions but also start infusing all other workloads. Over the longer term, every workload will have an AI component that may be inseparably integrated into an application. This means that more and more applications need to be trained, and retrained, using deep learning techniques. Thus we expect a significant increase in data and algorithms that will require corresponding infrastructure capabilities to run effectively and in a timely manner (often real time or near real time).

Closely related is our expectation that we are approaching the end of the homogeneous datacenter because various types of processors other than classic x86 are bridging the performance gap that AI applications have so starkly revealed. These other processors can be different CPUs or accelerators or a combination of the two.

CHALLENGES/OPPORTUNITIES

Challenge

- **Confusion reigns:** Businesses are unsure what AI use cases will bring them business benefits, what skill sets they need to bring AI capabilities in-house, what software they should develop these applications with, what the infrastructure and deployment model should be, and what accelerated technologies to choose from to overcome the limitations of today's server infrastructure.

Opportunity

- **Effective and efficient AI computing:** IDC believes that out of this chaotic environment, a model will emerge for effective and efficient AI computing — those vendors that work closely with customers to experiment, scale and, ultimately, bring AI capabilities to their entire business will develop the wherewithal to define the right AI models from a hardware, software, and deployment perspective and bring them to market. These vendors will be tomorrow's leaders in the AI computing space. Businesses should look to identify these emerging leaders for their AI initiatives.

CONCLUSION

IDC has observed that, for businesses that are getting started with AI and deep learning, a period of trial and error on existing hardware is very common. As deep learning algorithms and AI applications are being investigated, experimentation with server infrastructure to run these new workloads on should be encouraged, as should a stance from server vendors to — somewhat altruistically — help their customers with this complicated stage. At the same time, however, the infrastructure team needs to prepare for the next stage, when production will be initiated for the AI applications under development. AI and deep learning are extremely demanding on server infrastructure and benefit from specific configurations, CPU characteristics, I/O capabilities, accelerators, and interconnects between the CPU and the accelerators. For many AI initiatives, whether small or large, businesses will benefit from getting support, which can be solicited from various parties. Perhaps most effectively, however, such support is provided by a server vendor that offers a complete AI hardware/software stack that is wrapped in an all-encompassing support strategy all the way from the initial experimental stage to a well-integrated and scalable implementation of the AI solution.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com