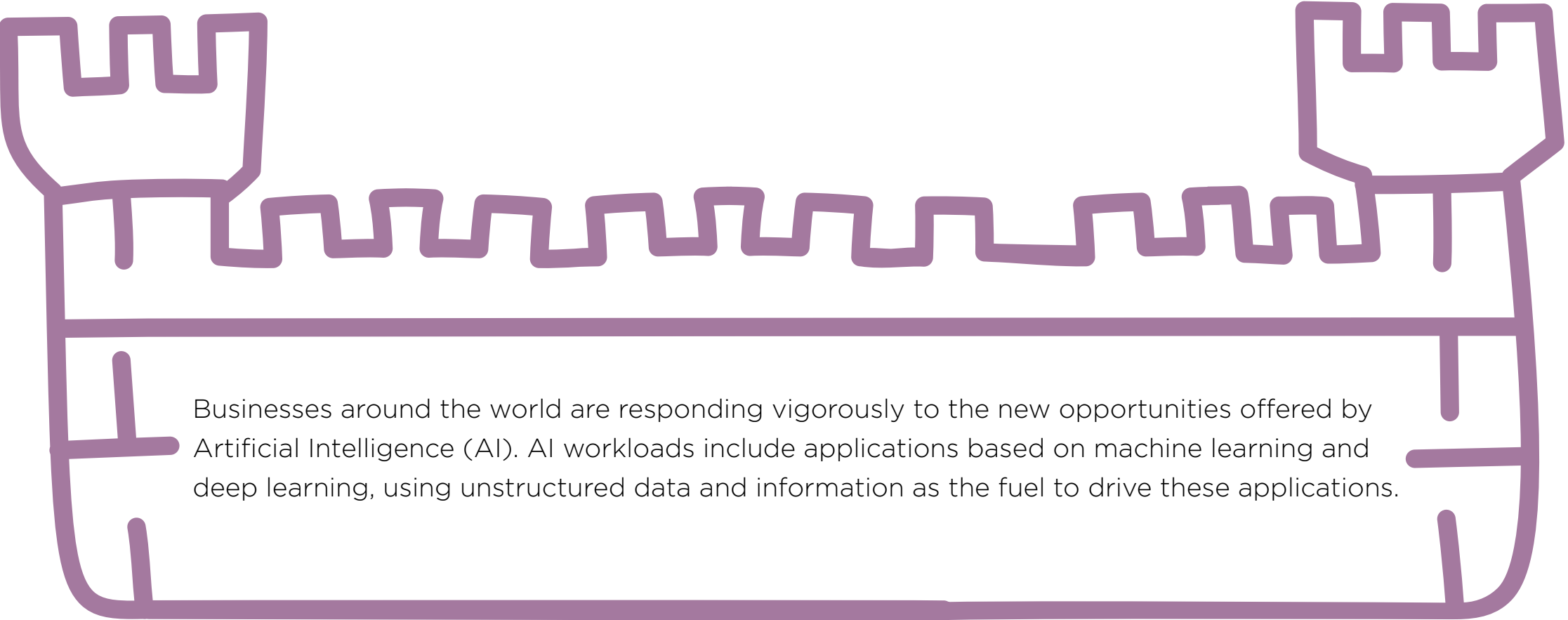




Considering IT Infrastructure in the AI Era

An IDC InfoBrief, sponsored by IBM | October 2017

Hitting the wall with infrastructure on your AI journey?



Businesses around the world are responding vigorously to the new opportunities offered by Artificial Intelligence (AI). AI workloads include applications based on machine learning and deep learning, using unstructured data and information as the fuel to drive these applications.

AI is taking off – in 24 months, **25%** of all workloads are expected to be AI.

Deep learning – which is extremely compute intensive, takes up to **50%-60%** of an AI workload.

AI workloads — including **deep learning** — demand specific and high-performing server infrastructure that many businesses are struggling to identify and build.

Businesses are trying all types of infrastructure for their AI workloads

Ranking from most used to least used

- 1.** A cluster of 1- or 2-socket servers (with accelerators)
- 2.** A cluster of 1- or 2-socket servers (no accelerators)
- 3.** A cluster of scale-up (4+ sockets) servers (with accelerators)
- 4.** A cluster of scale-up (4+ sockets) servers (no accelerators)
- 5.** A traditional dedicated server
- 6.** A high performance, high density solution
- 7.** A converged server (no accelerators)
- 8.** A packaged solution of server hardware and cognitive software from a third party (with accelerators)
- 9.** A packaged solution of server hardware and cognitive software from a third party (no accelerators)
- 10.** A hyperconverged server (with accelerators)
- 11.** One or more VMs or partitions on a virtualized server
- 12.** A hyperconverged server (no accelerators)

Yet they're hitting the wall with their AI infrastructure and generational shifts are happening fast, in all directions

Current generation of the AI infrastructure that businesses are on



1st

39.6%

2nd

37.6%

3rd

22.8%

Generational infrastructure shifts that businesses have gone through



Top 7

1. Greater processor performance
2. Scale-out to scale-up
3. VM to dedicated server
4. Scale-up to scale-out
5. Greater I/O bandwidth
6. Dedicated server to VM
7. Added accelerators

More than **45%** of small businesses and **35%** of large businesses expect their current infrastructure for AI to last **no more than another year.**

Indeed, **15%** are running into limitations today. 

In the next 24 months, the use of accelerators in infrastructure for AI will therefore grow significantly, including GPUs, FPGAs, ASICs, and Many-Core Processors.



Businesses are expecting a significant performance boost from these accelerators for a manageable price premium.

There will also be a distinct migration to the cloud for AI workloads

75% of businesses that expect to run AI ONLY in the cloud in 12 months are both on-premise and in clouds today. In other words: **their cloud experience has so far been satisfactory.**

However, this migration to the cloud will not affect the overall distribution of AI workloads between cloud and on-premise.

In 24 months, **45%** of businesses still expect to run AI on-premise and **23%** will run AI in the cloud.

In the cloud, businesses run into more limitations with compute for AI than on-premise

Top 10 Limitations in the Cloud

1. Manageability
2. Scalability
3. Performance
4. Completing tasks within SLAs
5. Storage
6. Diagnostics
7. Virtualization
8. Interoperability with the datacenter
9. Memory capacity
10. Load balancing

Yet on-premise has its own limitations

Given the challenges with AI computing in the cloud, accelerated compute on-premise is a valid choice. However, even with running AI on-premise there are challenges.

Top 5 Limitations On-Premise

1. Manageability
2. Performance
3. Energy use
4. Diagnostics
5. Completing tasks within SLAs

Acceleration alone is not the silver bullet – system and hardware platform architectural features, including core performance, I/O bandwidth, and manageability, matter just as much.

Apart from the hardware for AI, AI software and data management for deep learning are also complicating the journey to AI

28% find that the time to value with AI software is too long

25% can't manage data volumes with AI

23% don't know what the right software/algorithms would be for the challenge they're trying to address

23% have trouble keeping sensitive data for AI secure

22% have difficulties preparing data for AI

Recommendations for businesses on this AI journey

- » 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.
- » Don't just consider server products available from your traditional vendor, but look at other server vendors as well, especially those offering a complete AI hardware/software stack.
- » Some vendors provide support at all deployment stages of an AI system, from hardware selection and optimization through the software stack all the way to deployment and consulting services.
- » Select a vendor that has demonstrated thorough knowledge of infrastructure requirements for AI and deep learning.
- » Make sure the vendor can advise on the first experimental stages, even if that is on your existing hardware, and can then guide your organization toward on-premise or a hybrid on-premise/cloud expansion.
- » Choose a vendor that can work through various small, mid-size, and large AI scenarios so they can serve as an advisor for the small initiative but also as a consultant for a larger AI initiative.