When it comes to AI infrastructure, some of the pain points might surprise you

Analysts - Nick Patience, Jeremy Korn

Publication date: Wednesday, September 4 2019

Introduction

AI, machine learning, deep learning and all the other variants will place huge demands on IT infrastructure. Whether you are an enterprise with your own datacenter, a datacenter operator or some other form of cloud provider, organizations are becoming cognizant that AI is a resource hog, for reasons we will explore shortly.

Data compiled by OpenAI in 2018 shows that the compute being used to train the largest machine learning models – from AlexNet in 2012 to DeepMind's AlphaGo Zero in 2018 – was doubling every 3.5 months (considerably shorter than Moore's Law's doubling time of 18 months). This means that, from 2012-2018, compute resources used to train these models increased by more than 300,000 times. This is an abstract benchmark not specifically tailored to any specific use case.

However, now we have just such data in the form of our new Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019 survey, which we have just published. It shows that nearly half (48%) of responding enterprises indicate that their current AI infrastructure will not be able to meet future demands (see figure below). This means that the infrastructure demands created by AI and machine learning workloads will put pressure on IT managers – and the vendors that supply them – to integrate new products or re-architect existing systems.

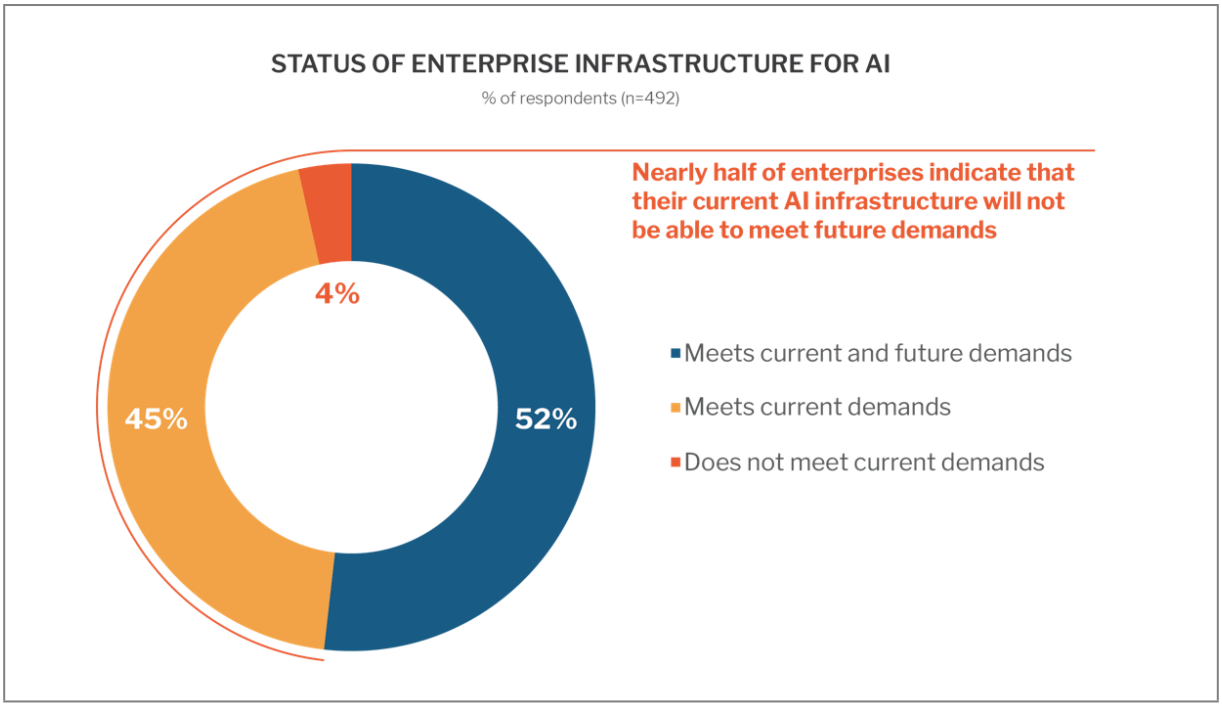
The 451 Take

The core of machine learning is mathematical computations that, for most enterprises, are unprecedented in both complexity and volume. These workloads increasingly require high-performance or application-specific infrastructure to ensure AI systems can be built and deployed quickly. The adoption of AI is causing enterprises and service providers to rethink their infrastructure needs, particularly in the areas of storage and compute resources, but also networking. Infrastructure vendors and cloud providers are only too happy to oblige with new dedicated resources for AI in the cloud and on-premises, but we're still in the stage of identifying where the

bottlenecks are and how to solve them. However, this isn't just about adding a few extra servers in a rack. In many cases, this is about a complete rethink from the level of silicon through to SaaS.

The status of enterprise infrastructure for AI

How capable is your organization's current IT environment of meeting expectations for AI/ML workloads?



Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019

Having a strategy for coping with these increased demands is already a minimum requirement, and the vast majority of organizations surveyed fulfill this basic criterion. Where they differ is how they will execute those strategies. Many are favoring cloud-based AI platforms, while on-premises and hybrid approaches also prevail. What all these approaches have in common is that resources are being dedicated to improving the performance of AI-based systems and applications.

Why the need for dedicated AI infrastructure?

AI is primarily driven by data. Data is used to build and train the models, and then new data is fed through the models to make inferences. How much data is needed to make this work depends on the type of AI used – classic machine learning algorithms such as Random Forest will use orders of magnitude less data than a 10-layer or deep neural network, for example. However, our survey shows that deep learning is indeed penetrating the enterprise, with nearly 50% saying they are already employing deep learning as part of their AI workloads.

The data volumes and sophistication of the algorithms are growing, but so is the amount of time being spent on data preparation and management, which our survey shows is the most resource-intensive part of the machine learning pipeline, ahead of machine learning model training and inferencing. Add to the mix a desire to use real-time data processing, versus batch, and you have yet another resource hog. We're not at that stage yet, but our survey says that is mainly because the current infrastructure doesn't support it, with about one-quarter of respondents saying they have no current interest in real-time data processing. Lack of performant AI infrastructure is also holding

back model retraining as often as organizations would like to – slightly over half of respondents said as much. The needs are there, but are not being met by current enterprise AI infrastructure.

What can be done, and will people pay for it?

As part of the aforementioned AI infrastructure strategies, enterprises are willing to spend money on new forms of dedicated AI infrastructure to address particular pain points. Some of those paint points might surprise you, such as higher-performance networking, which is cited by over one-third as the area enterprises would seek to improve to increase the performance of their AI infrastructure. This attests to an unspoken truth – the gains of implementing better compute and storage can be realized only if the connections between these components support faster data flows.

There is an appetite for AI-specific infrastructure, and our survey demonstrates that people are willing to pay for it. The features they are looking to spend money on vary significantly. For example, more than 80% express a willingness to pay for a single view of their entire data for their AI initiatives – even though such a thing might not be technically possible in large complex organizations. But along with the aforementioned higher-performance networking, enterprises are also looking to invest in accelerated cloud services, increased memory capacity, high-performance storage and hardware accelerators.