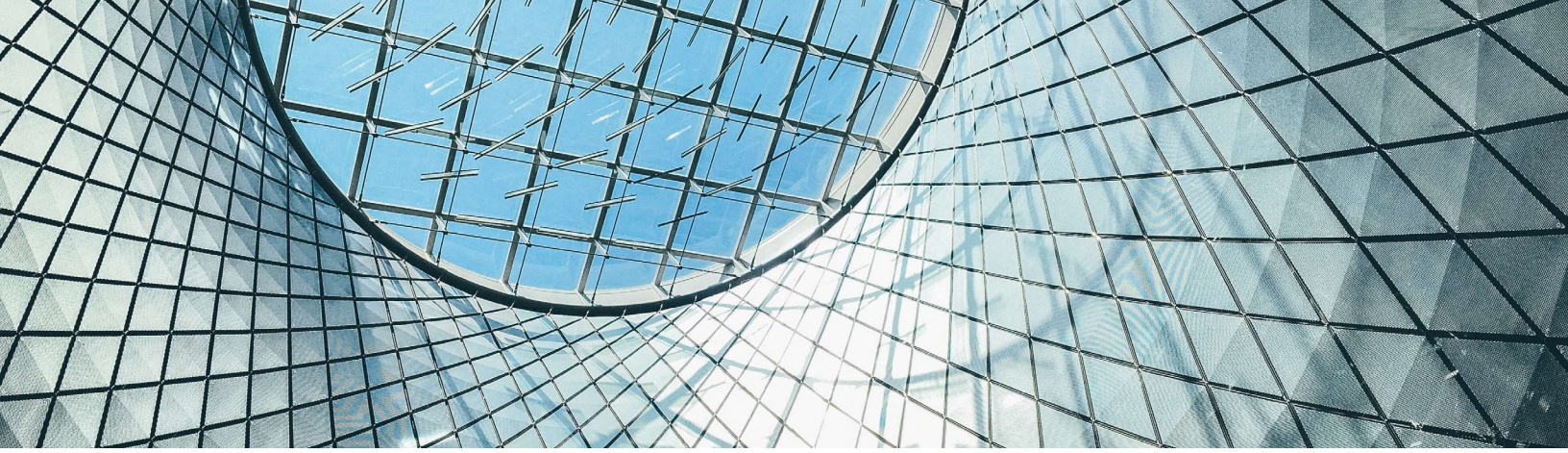

Market Insight Report Reprint

AI infrastructure is under strain, tested by growing workload demands – Highlights from VotE: AI & Machine Learning

July 19 2023

by Alex Johnston

Many organizations are struggling to meet the full scale of demand for AI capabilities. Discussion of this challenge has in recent years focused on GPU cost and availability, but the bottlenecks for organizations are significantly wider ranging.

S&P Global
Market Intelligence

This report, licensed to AMD, developed and as provided by S&P Global Market Intelligence (S&P), was published as part of S&P's syndicated market insight subscription service. It shall be owned in its entirety by S&P. This report is solely intended for use by the recipient and may not be reproduced or re-posted, in whole or in part, by the recipient without express permission from S&P.

Introduction

AI-centric workload investments drive growing demand for infrastructure, and most organizations continue to have a distinct strategy to address these workloads. High-performance infrastructure designed for AI — with processors, storage and networks optimized for machine-learning (ML) workloads — is required to efficiently take advantage of larger models, invest in lower-latency applications and leverage a greater diversity of datasets. 451 Research's Voice of the Enterprise: AI & Machine Learning, Infrastructure 2023 study reveals that many organizations are struggling to meet the full scale of demand for AI capabilities. Discussion of this challenge has in recent years focused on graphics processing unit cost and availability, but the bottlenecks for organizations are significantly wider ranging.

THE TAKE

AI workload requirements are outstripping the ability for businesses to service the wide array of ML projects and in-production capabilities spinning up across organizations. Targeted use cases for AI are diverse and broad, with many organizations making use of hundreds of models, and the vast majority expect workload requirements to only increase. In this environment of AI workload expansion, infrastructure is emerging as a critical bottleneck. Contrary to media coverage of this challenge, where the cost and availability of AI accelerators — GPUs in particular — are commonly presented as the totality of the issue, the infrastructure bottleneck is multifaceted. Higher-performance networking is as much a challenge as AI accelerators, and the cost of accelerators appears less a challenge than their reliability and performance.

Summary of findings

Infrastructure is the bottleneck as workload demands expand. IT infrastructure performance is the most common challenge that contributed to project abandonment at organizations over the past 12 months. Even for projects that successfully have made it into production, infrastructure appears to be a limit. Just 32% of respondents note their organization's IT environment was "always able to meet demand" for AI, while 67% believe they will need upgrades to their IT environment to meet future demand. Sixty-one percent of respondents noted that infrastructure limitations do, or will, prevent their organization from retraining models in production more often.

The workload demands placed on enterprises are significant, and likely to escalate further. The median average organization that has AI in production has 125 models and is using more than 1 PB of data to train those models in aggregate. Eighty-two percent of respondents expect AI/ML workload demand to increase, with just 5% predicting a decline.

Spending on infrastructure is increasing at an accelerated rate. Reflecting the barrier that infrastructure provides to organizations attempting to leverage AI, it is unsurprising that spending on infrastructure for AI/ML will increase for the vast majority (almost 90%) of respondents' enterprises. For a notable minority of respondents (16%), spending will increase by more than 50% — with a plurality of respondents predicting a moderate increase (25%-49%) in the next 12 months. These increases build on sizeable funds already invested in AI infrastructure, with 70% of respondents from organizations of more than 1,000 employees suggesting spending in the last 12 months exceeded $1 million, and 13% suggesting it exceeded $10 million.

The rate of increase for AI/ML infrastructure spending appears to be accelerating. Respondents in 2022 were mostly likely to only predict a "slight increase" over the next 12 months, rather than the "moderate increase" predicted by the 2023 cohort. Indeed, this "slight increase" appears to have been a conservative estimate, with a seemingly more pronounced change in spending than forecast. The proportion of respondents from organizations spending more than $1 million on AI infrastructure increased by almost 10 points from 2022 to 2023, with the median average increasing by $500,000 in the past 12 months to $1,250,000.

©Copyright 2023 S&P Global Market Intelligence. All rights reserved.

Accelerators for AI/ML training and inference are important, but so is higher-performance networking. There is a pronounced need for AI accelerators. Thirty-seven percent of respondents see a need for accelerators in the cloud to improve the performance of their workloads, 29% for on-premises GPU servers and 26% for stand-alone on-premises hardware accelerators. The need for such accelerators is growing — the figures for 2022 were 33%, 25% and 14%, respectively. Higher-performance networking remains the most commonly identified need, however, selected by 45% of respondents. Lower-latency networks with better communication bandwidths to ensure GPU processors are operating at full capacity are an area of focus for businesses, albeit one that receives significantly less attention.

Cost of hardware accelerators is less of an issue than reliability and performance. Inflationary pressure generated by AI accelerator shortages, particularly GPUs, is an issue. However, while businesses feel this pricing pressure, with 37% of respondents who use hardware accelerators seeing budget/pricing as a top concern, respondents are more likely to see reliability and performance as the most pronounced problems, followed by scalability. It appears that the importance of hardware accelerators, with faster computation and high-bandwidth memory, means that access to budget is less of an issue than for other areas of IT investment. Instead, the limitations of accelerators, be that performance for latency-sensitive applications or high in-field failure rates, are of greater concern to respondents. The challenge of scalability is perhaps unsurprising, coming not just from the general pressure from growing workload demands, but also from some accelerator-specific problems, with some designs such as application-specific integrated circuits trading reconfigurability in return for greater energy efficiency, for example.

Data-intensive workloads and regulatory compliance are driving AI to the edge. Edge computing plays an important role within AI/ML strategies, aiding compliance, supporting data-intensive workloads and enabling new applications. More than half (54%) of respondents see the edge as “extremely important” in maintaining regulatory compliance, and 53% in supporting data-intensive workloads. Just 4% of respondents disagree that their organization would prefer to conduct more training and inference in edge locations. Budget, a lack of skills or experience at the edge, and storage capacity are seen as three major bottlenecks to training and inference at the edge.

Security, reliability and cost are major concerns, but so is sustainability. Security, cost and reliability represent the three largest concerns organizations had around their AI/ML infrastructure. However, concerns around sustainability, which 31% of respondents suggest is a consideration for their organization, outstrip many long-standing focus areas, such as time to value, data sovereignty and manageability. Organizations also appear to be reacting to these concerns, with 63% of respondents from organizations using public cloud infrastructure suggesting sustainability was not only a factor in choosing a region or availability zone, but that they made a decision because of it. The majority of survey respondents suggest their organization would be willing to pay to achieve more sustainable AI/ML infrastructure, with around one-third noting that it is essential to their organization and that they would pay a premium for it.

©Copyright 2023 S&P Global Market Intelligence. All rights reserved.

CONTACTS

The Americas
+1 877 863 1306
market.intelligence@spglobal.com

Europe, Middle East & Africa
+44 20 7176 1234
market.intelligence@spglobal.com

Asia-Pacific
+852 2533 3565
market.intelligence@spglobal.com

www.spglobal.com/marketintelligence

Copyright © 2023 by S&P Global Market Intelligence, a division of S&P Global Inc. All rights reserved.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively, S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers, (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON "AS IS" BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT'S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence's opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global Market Intelligence does not endorse companies, technologies, products, services, or solutions.

S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain non-public information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its websites, www.standardandpoors.com (free of charge) and www.ratingsdirect.com (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at www.standardandpoors.com/usratingsfees.