ROBERT FRANCES GROUP
Business Advisors to IT Executives

May 25, 2022

Summary Insights: Meeting Growth Amid Hardware Shortages and Energy Constraints

RFG Perspective: Satisfying enterprise growth and transformational requirements in data centers in 2022 is a major challenge for IT executives. Near-term capacity problems, constrained supply, and rising prices are impeding IT executives' ability to deliver additional data center capacity and to contain costs globally.

While one can order new server, storage, networking and related equipment without issue, vendor backlogs are not in alignment with business timeframes – and this is forcing hard decisions. Supply-chain issues for microprocessors and components are causing delays of up to six, nine, or even 24 months for many types of vendor hardware, RFG customer surveys show. And it's clear that challenges related to availability sourcing, manufacturing, and shipping are intensifying for many vendors this year.

Similarly, a variety of environmental issues, including energy scarcity; rising costs related to demand growth; carbon-neutral and net-zero environmental initiatives, and global tensions – all are impacting data centers' power demands. To enable their strategic growth requirements, enterprises will need to balance existing hardware and limited new hardware acquisition capabilities with limited power supply constraints. In 2022, IT executives will need to find approaches that will utilize existing resources more efficiently and sustainably, so that they can satisfy business demand through 2025.

INTRODUCTION

While the IT challenge presents itself primarily in the context of growth and capacity, and sustainability initiatives imposed by regulatory bodies, other factors must be considered, including the support of CEOs, internal ESG organizations, and corporate stakeholders.

More than 80 percent of S&P 500 companies publicly report sustainability metrics in some form – one component of which is their IT sustainability progress. In addition, supply chain constraints will likely remain for at least the next 18-24 months for new hardware and power supplies, causing user costs to increase between 30 percent and 400 percent, across a range of geographies worldwide.

Discussions incorporating the board, C-level executives, line of business executive, and IT executives must deal with these new realities, prioritize desired and essential outcomes, and find ways to resolve their business issues. They also must satisfy their company's stated ESG commitments more efficiently.

One key indicator of efficiency is the PUE (power usage effectiveness) ratio. Along with other energy-consumption, efficiency-oriented metrics, PUE data will need to be added to the price/performance evaluation criteria. What is needed is a 360-degree examination of all of these factors, incorporating key metrics for facilities, IT hardware, and power and cooling.

Existing strategies must be evaluated, planned for, addressed, and routinely measured, before identifying ways to optimize and maximize efficiencies, thereby minimizing the effects of supply chain shortages. There are a wide variety of options to optimize and maximize data center efficiencies, as discussed below.

The Panel:

The Robert Frances Group hosted a videoconference on April 20, 2022, to discuss the capacity challenges and options that data center IT executives must consider when deciding where to deploy workloads – locally, or on the cloud. Panelists included:

- Jen Huffstetler, VP, Intel
- Sarah Grzybowski, Portfolio Strategy Lead, IBM
- Travis Wright, VP, QTS Data Centers
- Rob Bunger, Program Director, CTO office, Schneider Electric

Exploring Strategies for Improved Data Center Efficiency

The panelists and participants identified the following areas for improvement:

Data Center Hardware

Over the next few years, demand-based growth in the data center is projected to average at least eight to 10 percent on the service side – rising to 20 percent or more on the storage side. To meet the increased demand, IT executives need to find a way to maximize the assets that have already been installed.

The good news is that, in most cases, server, storage, and networking hardware are not maxed out; the bad news is that correctly identifying all of the underutilized assets and implementation of adjustments is not a simple task. Many customers are applying monitoring and AI-based software tools to identify applications where energy use and efficiency could be optimized to save money – and improve business outcomes.

Server utilization industry averages are still embarrassingly low. One reason is that the systems were originally over-provisioned. The DevOps or the development teams that make the decision concerning server usage storage requirements may overestimate the requirements – either to avoid overloads during peak periods or because of overestimating needed capacity – resulting in over-provisioned hardware.

Re-balancing Hardware Resources

One key "lever" that can be adjusted in the data center: the ratio of memory to CPU usage requirements is often unbalanced. That creates an opportunity for IT organizations to rebalance their hardware resources, or to redesign their software for better overall alignment of software to hardware – resulting in better utilization of a company's on-premises systems.

Hardware efficiencies can be improved by increasing virtualization, containerization, port capacity, storage usage, as well as by improving code density. While some fixes may be easy to do, others are not. For example, by using virtualization, customers can achieve 30 percent server utilization with limited infrastructure or code changes. But to drive that number to higher levels of utilization, other steps – such as code rewrites, re-platforming or switching to containerization – may be necessary.

Clearly, a powerful option is to migrate selected workloads to run on the hybrid cloud – including public and private (dedicated) clouds.

During the COVID-19 pandemic period, the pace of cloud migrations has been accelerating. Customers are adapting and modifying data center applications so that scale-up applications can run on cloud infrastructure. If cloud migration for scaling up these demanding workloads is the goal, then the application's security, resiliency and availability must be carefully considered, along with any dependencies related to using a data center's on-premises systems. Finally, organizations must test the modified applications to be sure that they are making the best, most efficient use of on-prem resources (in the data center) and off-prem resources (in the cloud).

Leveraging the cloud's capacity for scaling up "bursting" workloads, on demand, is a good reason to migrate a given application from on-premises hardware.

Leveraging AI-based software is another way to improve infrastructure utilization. AI tools and automation make it possible to dynamically move workloads as workloads peak and drop throughout the day, based on usage-driven demand for these workloads.

Similarly, IT organizations need ways to prioritize application availability, based on business need. That way, they can automatically "throttle down" certain lower priority applications or functions during peak processing times (e.g., according to time-of-day, or to when demand rises during a holiday-related sales promotion).

Improving Software Efficiency with AI Tools

Software efficiency is another area for improvement. It is estimated that software efficiency can be improved by 35 percent by eliminating inefficient code. The RFG 100 panel suggested that one approach is for IT executives to make sure that they are using the latest updates to their site's system software and the latest drivers for hardware functionality.

Many software solutions have built-in monitoring capabilities that should be utilized to optimize operational efficiency. Additionally, there may be built-in AI features in software development code that will allow DevOps to inspect and repair aging code. In many cases, software code efficiency can be improved by using AI tools, industry-standard APIs, and by using low-code/no-code programming techniques to refactor existing applications.

In some use-cases, blockchain should be avoided, because implementations can be inefficient – leading to over-consumption of power and cooling resources. For example, one enterprise required 600 virtual machines (VMs) just to process a few thousand transactions a day. That is not energy-efficient or resource-efficient. If IT executives are deciding to build new solutions to improve data-processing efficiency, then choosing energy-efficient hardware and more efficient software programming should both be a top-priority consideration.

Capacity Planning

Capacity planning for many organizations has become a lost art. In many cases, companies buy additional hardware to handle increased workloads, or they move the workloads to the Cloud or the Edge. This is especially true for racks and racks of scale-out x86-based hardware systems, which have surrounded traditional scalable systems over time.

Configurations should be checked to ensure efficient deployments – or to consider moving applications and databases to other configurations. However, capacity planning reviews have long been part of continuing operations for scalable mainframe systems – and scalable systems often support built-in capacity-on-demand features that add processors, as needed.

Inside the data center, using discovery tools, DCIM (Data Center Modernization software), and resource monitoring dashboards help to determine which physical assets are degrading or malfunctioning. Once shortfalls are detected – either though routine monitoring or as part of scheduled maintenance – then Operations teams should work to improve operational efficiency by adjusting the way software is deployed within the data center.

Leveraging PUE Metrics

Counterintuitively, power Usage Effectiveness (PUE) can become worse over time before it gets better – as IT equipment becomes more efficient, even if power and cooling elements remain unchanged. Nonetheless, PUE should be one of the key metrics of customers' energy-efficiency and ESG efforts. [We note here that the PUE ratio is calculated by dividing the total data center facility power used by the IT hardware power consumed].

There are significant differences in PUE between those of cloud providers, those of colocation providers and those of most large enterprise data centers. Most cloud and colocation providers have a PUE between 1.3 and 1.05. By contrast, most enterprise data centers are running with a PUE of 1.6 or worse – and that could, and should, be improved, wherever possible.

It is not unusual to find enterprise data centers that have PUEs greater than 2.0 – which says that for every $1 (USD) spent on running the IT equipment, there is an additional dollar spent on the supporting facility's infrastructure. In 2020, 43 percent of a data center's electricity usage, on average, was spent on power and cooling. Cooling is the largest overhead component – with the powertrain coming in as the second largest overhead component.

Powering Data Center Sites for Growth

At one time, new data sites were planned to operate at a maximum of 30 MW, but many new sites are being planned now for 200 MW – with some sites planning on growing to 1 GW. While it is possible to go from an empty lot to an operational data center in just 11 months of construction, it could take up to two years for a power company to deliver enough electrical power required to run that data center.

Given current supply constraints, data center growth is turning out to be a major challenge. Innovative approaches, like re-purposing other on-campus buildings, can be used to "grow" data center real-estate, without constructing new facilities. Some other brownfield reuse approaches are re-purposing aging semiconductor plants, or old distribution warehouses into new data centers. Others are consolidating multiple data centers into a smaller number of energy- efficient data centers.

Some companies are exploring the option to switch to alternative energy supplies, such as solar power or wind-power. But switching energy sources is not a short-term fix. Delivery of some brand-new solar energy systems may be delayed by a year, or more, due to supply-chain issues.

Two Use-Case Examples

1. **Modular Pods:**

Modularization can impact a data center's energy envelope. Use of preconfigured and fully-loaded modular pods (usually 20' x 40'), delivered in 1.5 megawatt (MW) chunks, is a quick and efficient way to grow a data center. The pods use low-pressure refrigeration instead of being water-cooled. These pods can be designed to have a PUE of 1.25 from the start, in contrast to traditional designs that have a PUE greater than 3.0, but do not achieve a lower PUE until they are fully populated. One more benefit of this approach: modular pods can be pre-ordered and kept in storage (rather than planning to use just-in-time (JIT) delivery of equipment.

2. **Tapping Renewable Energy:**

For enterprises that have multiple data center sites, IT executives should consider emulating what hyperscalers (cloud service providers, or CSPs) are doing to address their power issues.

Hyperscalers monitor the amount of renewable energy that is coming onto the power grid at any given time. They can shift their workloads if the amount of power provided by solar-driven or wind-driven energy sources are reduced, due to changing weather conditions. Even when weather conditions change, this kind of flexibility in tapping different kinds of power/cooling capabilities has two advantages: it results in improved carbon efficiency and it improves business resiliency, as well.

Heat-Mapping Data Center Resources

To best understand existing heat loads inside a data center, IT and facility executives should consider using thermal (or fog-based) heat maps. This approach to outfitting data centers will help them understand where their compute loads are generating more heat within their space. Most importantly, it will allow them to balance "hot" and "cold" spaces in a much more thoughtful way. It will reveal high heat loads being generated by systems, storage, and aging UPS systems. It is worth noting that heat-mapping tools are now far less expensive – and much more portable – than was the case just a few years ago.

CFD Tools:

For purposes of cooling, IT organizations could consider using CFD (computational fluid dynamics) tools when IT organizations are planning design changes for new or repurposed data centers. CFD tools allow planners to look at the air flows inside the data centers, allowing facilities designers to model changes before a data center build-out ever takes place. This approach can result in significant savings, because in many data centers, cooling units may be 10 to 20 years old, resulting in inefficient cooling systems, compared to current models.

Automation and Optimization

Automation and optimization should be the key directives for a data center's ongoing Day 2 operations. To optimize data center operations, IT organizations will need the ability to dynamically move workloads across enterprise IT infrastructure without human intervention. This is becoming more urgent, as companies move to hybrid cloud and multi-cloud deployments.

We believe that optimization should be done on a continuous basis, leveraging AI software tools to do so. Changes in environmental factors affecting the data center should be done on an hour-by-hour basis – given shifting levels of energy, carbon emissions, and water usage that occur throughout each 24-hour daily cycle.

In aging data centers, monitoring and controlling humidity and temperature levels can be challenges for many IT organizations and facilities managers. Major factors in high energy consumption include: data transmission, GPU workloads, and the use of serverless computing. While applications can be moved to the cloud, via containers and Kubernetes, data should be kept as local as possible for purposes of efficiency, high performance, and compliance with regional governmental regulations.

Data transfers can be lengthy, due to network latency. This causes noticeable differences in overall system performance, especially for workloads and analytics that are accessing multiple data sources . Tapping GPU resources in the cloud, for example, reduces heat build-ups inside the data center. There are substantial GPU cluster resources available from cloud service providers. So, some HPC workloads would become more efficient if migrated to cloud providers who support intensive HPC applications. Otherwise, customers would be faced with the task of building up and rebalancing large and growing GPU clusters that are used for high-performance computing (HPC) and data analytics.

SUMMARY

In an era of supply-chain issues, climbing energy prices and cloud migrations, modern data centers must focus on operational efficiency and environmental sustainability, and growth.

That's why IT and facilities executives must focus on improving the energy profile and the capacity of a data center's on-premises server and storage systems. While cloud computing is reducing the need to build out local data centers, the data centers that remain in use must develop strategies for deploying high-efficiency sustainable solutions, as outlined in this RFG 100 panel discussion.

RFG POV: Executives need to take a holistic approach to addressing efficiency, sustainability, and growth in their data centers. This will require an investment in resources and skills – and possibly new processes – rather than focusing on the acquisition of new equipment alone.

Given current supply chain constraints, senior corporate management -- including business, IT, and facilities executives – needs to invest now to improve the efficiency and sustainability of their operations. That will allow their enterprises – and the business units within them – to meet their business requirements while optimizing their power/energy profile.

Additional relevant research and consulting services are available. Interested readers should contact Client Services to arrange further discussion or interview with Cal Braunstein, CEO and Executive Director of Research.