ROBERT FRANCES GROUP
Business Advisors to IT Executives

Oct. 30, 2024

Summary Insights: 10 Ways to Cut Data Center Costs

RFG Perspective: IT executives should be able to reduce data center costs by more than 40 percent, which should help them fund investments that better enable current applications, develop new generative AI (genAI) apps or other solutions that improve the enterprise's competitive advantage. Most data centers today are not monitoring and measuring their operations' performance so that they can run at optimum or maximum efficiency. There are 10 levers at their disposal that they can use to lower data center costs short- and long-term whether the workloads are in on-premises data centers, colocation sites, or in the cloud. While some of the benefits may go to the Facilities group, most of the savings will accrue to the IT operations group. In that the demand for more compute and solutions is increasing and that power is, and will be constrained over the foreseeable future, the time to act is now.

INTRODUCTION

Most IT data center executives use a CapEx model for their method of hardware acquisition. This approach requires less effort and less personal risk for IT executives. But it is the least desirable means of funding data center transformations. It falls short for two reasons: cloud and colocation provider purchasing use forms of the OpEx model and are easier to gain buy-in and execute; and it reduces flexibility. One advantage of the CapEx model is that IT executives do not require a good asset management system in place (a common data center shortcoming), which is a requirement for the OpEx models if one hopes to keep costs under control.

On the other hand, while the OpEx models require a better understanding of financial analysis and cost structures, it allows for greater flexibility and reduces operational risk. There are multiple OpEx models to choose available. Cloud service providers (CSPs), colocation providers and hardware suppliers offer multiple OpEx versions, each has its pros and cons.

Rack Density and Other Facilities Concerns

IT lifecycles are much shorter than the facilities lifecycles. Facilities plan and build on the assumption that the data center will be operational ten to twenty years or more. IT thinks in one-, three-, and five-year terms. This disconnect has long-term implications. Most data center rack, aisle, power, and HVAC buildouts are Facilities concerns and they design for the long-term – or at least their vision of how that will occur. As a result, data centers and hyperscaler facilities have different architectural requirements and design points.

The introduction of genAI applications is wreaking havoc on data center facilities. Most data center racks are designed with less than 20 kW of power – with many having less than 10 kW. GenAI systems require much more power. Today a fully configured system of Nvidia servers on a rack requires a minimum of 50 kW, with the newer, more dense ones consuming over 100 kW. Making room for genAI systems may require Facilities to refit the data center and for IT to modernize their systems so that the new genAI systems can operate within the existing power and space envelopes. Moreover, the added power demand also results in an accompanying demand for cooling. This could be air cooling, or more likely with genAI, some form of liquid cooling.

Proper data center design and layouts can result in up to 30 percent savings alone. People forget that the data center is built to last for more than 20 years while the IT equipment gets upgraded or expanded yearly. These changes can impact the air flow, cooling, and internal fan usage, and create hot spots that drive up costs. Often IT makes the changes without input from Facilities or without performing air flow or hot spot analyses. Working more closely together, can enable the executives to be more efficient and lower the costs.

Leasing and the OpEx Models

Colocation providers operate using both retail and wholesale models. In the wholesale model, which they use when providing space for hyperscalers and some large enterprises, pricing is by kilowatt. The retail model is used with data center executives who want to talk about square footage. Retail buyers will end up paying for space and power. Most data center executives lease the colocation space but still buy the equipment that fills the space. By utilizing a colocation provider, IT gets flexibility for growth but by purchasing the IT equipment, IT constrains its growth opportunities. IT executives should switch to an all OpEx model, which would be easier to fund and budget on an ongoing basis.

Many IT executives are leery of acquiring systems on an OpEx basis. One way to address this is to bring in a vendor or independent financing arm to acquire the systems and lease them back. If this is pursued, IT executives can lump into the purchase the cost of all the associated hardware and software licenses, and even hardware and software maintenance (for a three- or five-year period, or if possible, perpetually). Purchasing all these components up front and leasing them back will be less costly than buying them annually, as one can avoid any annual pricing uplifts.

If the IT organization is planning on executing a private cloud model, then an OpEx model is almost a must. In this way, expenses align with revenues (i.e., IT executives should be using a chargeback or showback model to the lines of business and corporate business units). This model will help with the implementation costs of genAI systems and will put the onus for the full payment of these systems to the business units underwriting the systems.

(Did You Know: doing a networking-only analysis for a data center rollout could result in significantly higher operational costs? By combining and performing a networking and compute

demand analysis simultaneously, an enterprise could reduce the overall total networking cost by up to 40% or more. Converged, blade-style compute architectures require up to 75 percent fewer network ports, cables, and cable runs allowing for faster deployment times and reduced implementation and support costs. With data center growth bandwidth averaging 12 percent per year (AI bandwidth doubling yearly) and growing, a converged compute environment radically contains the network device and rack space sprawl needed to enable growth and digital transformation strategies while maximizing efficiency in cost, power, space, sustainability and agility.)

AI

The initial CapEx investment in AI is typically significant, with the development cycle likely exceeding one year. That is why it is essential to plan early and thoroughly, acknowledging that AI deployment is not a quick-fix solution but rather a strategic, long-term investment. GenAI planning should involve development, facilities, and operations at a minimum. Each of these team should be developing a project plan that has dependencies upon the other; so, it is important that these plans be developed simultaneously. In this way each team will understand the expected bottlenecks and plan accordingly.

The development team will focus on the applications, data models, and data needed to satisfy the business requirements. The operations team will have to determine how they can deliver on the projected power and space requirements for the inference and training systems. For most companies, the inferencing systems could be deployed onsite or in a colocation site. The training systems may end up being in the cloud or in a hybrid mix. The plan will likely result in a data center modernization effort and/or an approach to satisfy the business needs through an onsite data center, colocation sites, and/or cloud services.

The facilities staff will have to determine how they can satisfy the genAI power and space requirements within the existing power and space envelopes. For many companies, it may require part of the modernization effort to have occurred before there is sufficient room for the building upgrade that is needed to support the new genAI system loads.

IT executives should be looking beyond just the initial install. GenAI systems have a tendency to grow – requiring additional power and space. Moreover, each generation of solution has been driving up the power density per rack. In addition, if the project is successful, there will likely be the desire to add additional genAI solutions for other parts of the organization.

Corporate Alignment

Most major enterprises today have made sustainability commitments. IT has tended to be the last group to engage in addressing the issue – aside from reporting its current status. IT could add a set of metrics to its current efforts, which shows its support of sustainability by driving operational efficiencies. The areas that could be monitored and improved upon over time are:

- Carbon emissions
- Circular economy
- Energy (renewable, and power usage effectiveness (PUE))
- Data center equipment
- Rack density
- Software
- Space
- staffing
- Supply chain
- Water
- Waste

SUMMARY

IT executives have been struggling with providing "more for less" every year for more than a decade. The twin demands of sustainability and genAI are pulling IT in different directions. But that need not be the case. If done judiciously, IT executives can simultaneously meet both these needs while also satisfying the demands of the business.

RFG POV: IT executives must find a way to meet current and future business demands, including new genAI applications, within budgetary, power, and space requirements. For many IT executives, this will require a change in their business models and require them to work more closely with various parts of the organization so that projects do not succumb to cross-group bottlenecks or challenges, or to extensive cost overruns. This will require detailed interorganizational planning, development, implementation, and monitoring of new measurements, and operating the data centers at optimum efficiency. If IT can satisfy the various crosscurrents, it can reduce data center costs by up to 40 percent or possibly even more.

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.