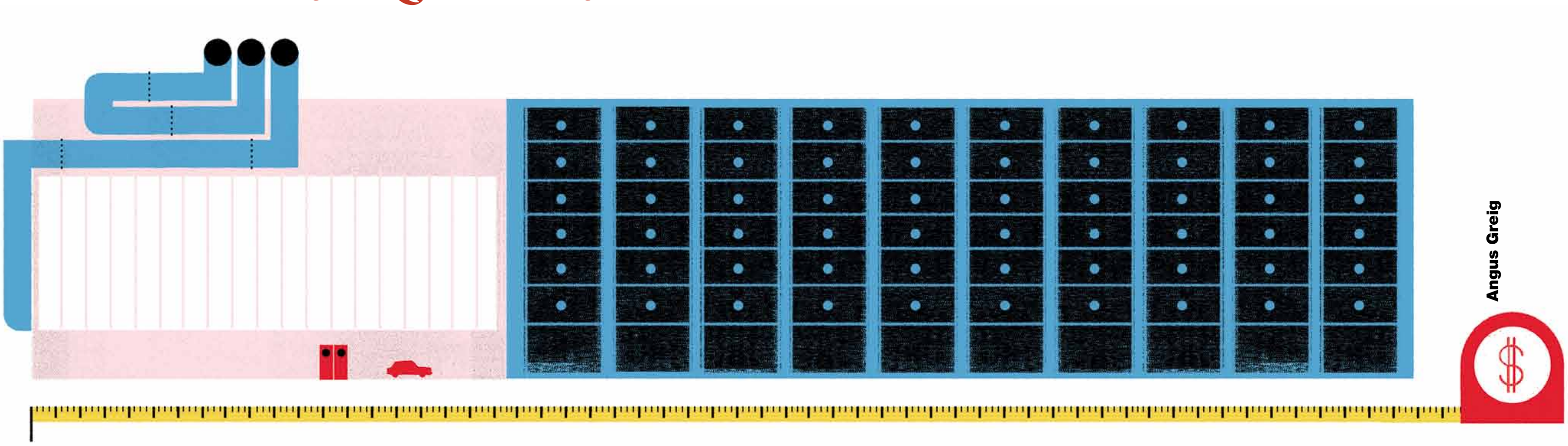
DECEMBER 2011

McKinsey Quarterly



Angus Greig

BUSINESS TECHNOLOGY OFFICE

Sharpening data center due diligence

CIOs should ask six questions before going to the board with a capital request.

James Kaplan,
Brent Smolinski, and
Allen Weinberg

Asking a board of directors for several hundred million dollars to obtain new data center capacity is one of the least popular requests a senior technology executive can make. As one CIO said, “I have to go to the executive committee and tell them that I need a billion dollars, and in return I’m going to give them exactly nothing in new functionality—I’m going to allow them to stay in business. I’m not looking forward to this.”

Investments in data center capacity are a fact of business life. Businesses require new applications to interact with customers, manage supply chains, process transactions, and analyze market trends. Those applications and the data they use must be hosted in secure, mission-critical facilities. To date, the largest enterprises have needed their own facilities for their most important applications and data.

How much data center capacity you need and when you need it, however, depends not only on the underlying growth of the business but also on a range of decisions about business projects, application architectures, and system designs spread out across many organizations that don’t always take data center capital into account. As a result, it’s easy to build too much or to build the wrong type of capacity. To avoid that, CIOs should ask a set of questions as part of their due diligence on data center–investment programs before going to the executive committee and the board with a capital request.

1. How much impact do our facilities have on the availability of important business applications?

Resiliency is among the most common justifications for data center investments. Out-of-date, low-quality data centers are often an unacceptable business risk. Yet upgrading facilities typically isn't always the most direct or even the most effective means of making applications more available. At the margin, investments in improved system designs and operations may yield better returns than investments in physical facilities.

Downtime overwhelmingly stems from application and system failures, not facility outages. An online service provider, for example, found that facility outages accounted for about 1 percent of total downtime. Even the most aggressive improvements in facility uptimes would have a marginal impact on application downtimes.

Organizations with high-performing problem-management capabilities can achieve measurably better quality levels by identifying and eliminating the root causes of incidents across the technology stack. Yet many infrastructure organizations do not have integrated problem-management teams.

2. How much more capacity could we get from existing facilities?

In many cases, older data centers are constrained by cooling capacity, even more than by power capacity: insufficient air-conditioning infrastructure limits the amount of server, storage, and network equipment that can be placed in these sites. The data center team can often free up capacity by improving their cooling efficiency, sometimes through inexpensive and quick-to-implement moves.

A large European insurance company, for example, wanted to consolidate part of its data center portfolio in its largest, most resilient data center, which was cooling constrained. The company freed up one to two critical megawatts of capacity in this facility—with approximately $40 million in capital cost savings—by replacing worn floor tiles, cable brushes, and blanking plates (all of which improved air flow) and increasing the operating-temperature range. As a result, the company consolidated facilities and provided capacity for business growth without having to build new capacity.[1]

3. What does future demand for data center capacity look like and how can virtualization affect it?

World-class data center organizations deeply understand potential demand scenarios. Rather than make straight-line estimates based on historical growth, they use input from business and application-development groups to approximate the likely demand for different types of workloads. They then model potential variations from expected demand, factoring in uncertainties in business growth, application-development decisions, and infrastructure platform choices.

Without a business-driven demand forecast, IT organizations tend to build "just in case" capacity because few data center managers want to be caught short. A large European enterprise, for

[1] Data center capacity is measured by the amount of electricity that servers, storage devices, and other equipment consume.

instance, cut its expansion plans to 15 critical megawatts, from 30, after the data center team conducted a deeper dive with business “owners” to better understand demand growth.

Even after years of rationalization, consolidation, and virtualization, many technology assets run at very low utilization rates, and every incremental server, storage frame, and router takes up space in a data center. Mandating that applications be migrated onto virtualized platforms in the facility, rather than moved onto similarly configured infrastructure, can be a powerful lever not only for reducing IT capital spending broadly but also for limiting new data center capacity requirements. A global bank, for example, cut their six-year demand to nearly 40 megawatts, from 57—a more than 25 percent reduction—by leveraging its data center build program to accelerate the use of virtual machines (Exhibit 1). This translated to a 25 percent reduction in new capacity build. That achievement helped create a political consensus for implementing virtualization technology more aggressively.

4. How can we improve capacity allocation by tier?

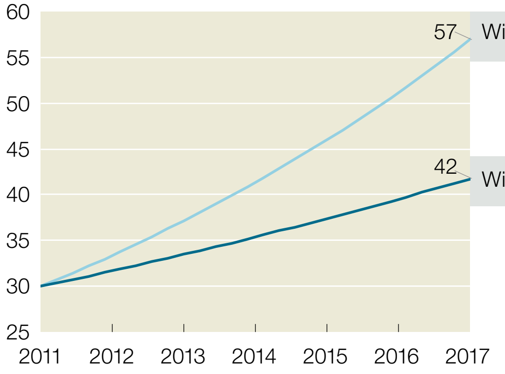
Owners of applications often argue that they must run in Tier III or Tier IV data centers to meet business expectations for resiliency.[2] Businesses can, however, put large and increasing parts of their application environments in lower-tier facilities, saving as much as 10 to 20 percent on capital costs by moving from Tier IV to Tier III

[2] The Uptime Institute has created four standard classifications for data center resiliency and functionality. Tier IV guaranties the highest level of reliability—99.995 percent uptime performance—by deploying multiple electrical and cooling backup systems. Tiers III, II, and I are categorized by diminishing levels of backup and redundancy, and thus lesser levels of performance. Tier I data centers guarantee 99.671 percent reliability.

Exhibit 1

Migrating applications to virtualized platforms can be a powerful way to limit data center capacity requirements.

Global bank’s data center demand and strategic capacity, megawatts



Virtualization	**Annualized growth,** 2011–17	**Virtual machines added,** 2011–17
Without	11%	0
With	6%	25,000

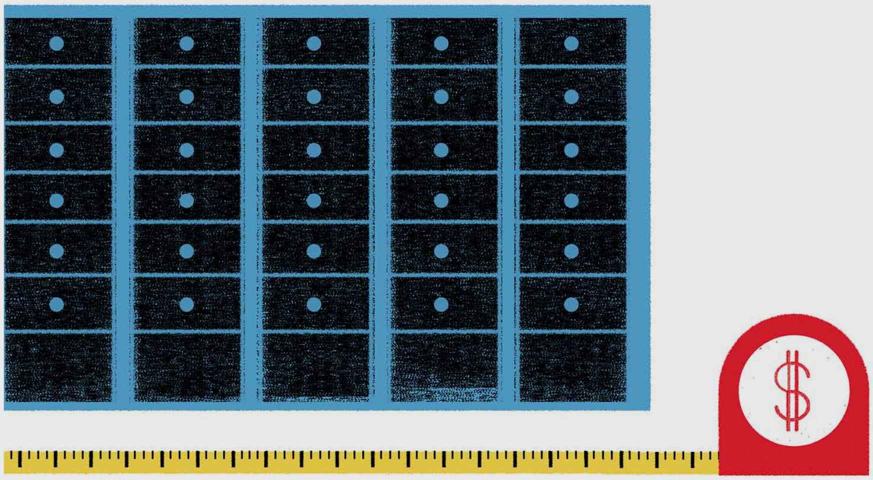
Three modular construction approaches

In-space cooling units
Specialized racks that can have self-contained cooling and electrical (for instance, UPS[1]) systems. In-space cooling units can be deployed in very small increments (rack size) and don't require a raised floor, so they can be deployed in almost any office space.

Containers
Prefabricated containers (about the size of trucking containers) that come prepopulated with server, storage, and network equipment and with cooling and electrical systems. They can be shipped to a facility and "plugged in" to a utility and network spline or fitting, require minimal construction within facilities (for instance, there's no need for raised floor space), and provide for small incremental capacity.

Modular buildings
Complete facilities, sometimes with raised floors, that are built in small, phased increments. These construction techniques often leverage prefabricated facilities. Extreme examples modularize mechanical and electrical plants, providing for varying types of plug-and-play technology.

[1] Uninterruptible power supply.



capacity (Exhibit 2). By moving from Tier IV to Tier II, they can cut capital costs by as much as 50 percent.

Many types of existing workloads, such as development-and-testing environments and less critical applications, can be placed in lower-tier facilities with negligible business impact. Lower-tier facilities can host even production environments for critical applications if they use virtualized failover—where redundant capacity kicks in automatically—and the loss of session data is acceptable, as it is for internal e-mail platforms.

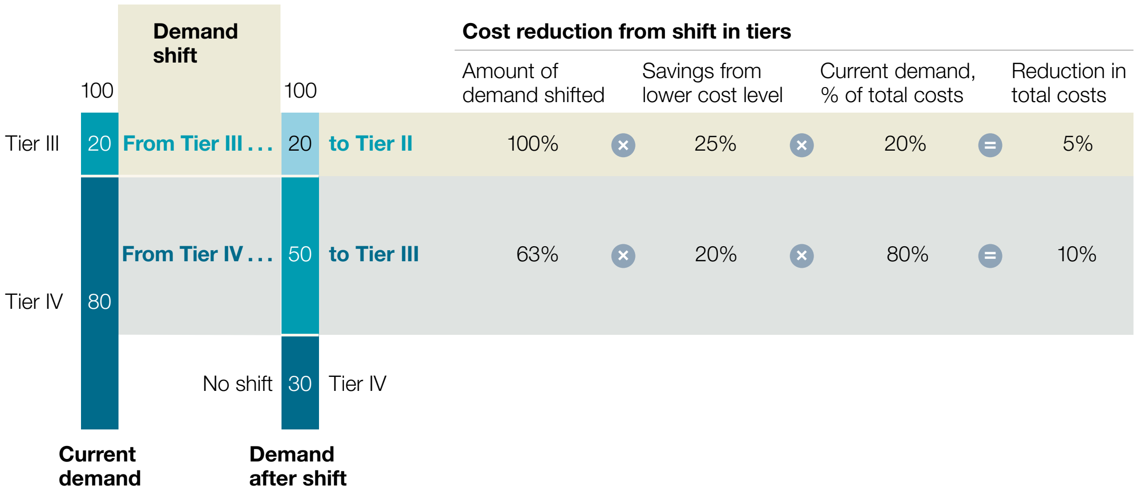
With appropriate maintenance, downtime for lower-tier facilities can be much less common than conventional wisdom would have it. One major online service provider, for instance, has hosted all its critical applications in Tier III facilities for 20 years, without a single facility outage. This level of performance far exceeds the conventional Tier III standard, which assumes 1.6 hours of unplanned downtime a year. The company achieved its remarkable record through the quarterly testing and repair of mechanical and electrical equipment, the preemptive replacement of aging components, and well-defined maintenance procedures to minimize outages that result from human error.

It is inherently more efficient and effective to provide resiliency at the application level than at the infrastructure or facility level. Many institutions are rearchitecting applications over time to be "geo-resilient," so that they run seamlessly across data center locations. In this case, two Tier II facilities can provide a higher level of resiliency at lower cost than a single Tier IV facility. This would allow even an enterprise's most critical applications to be hosted in lower-tier facilities.

Exhibit 2

Enterprises can reduce capital costs by moving large portions of their application environments into lower-tier facilities.

Proposed capacity build by tier,[1] %



Cost reduction from shift in tiers

	Amount of demand shifted		Savings from lower cost level		Current demand, % of total costs		Reduction in total costs
From Tier III . . . to Tier II	100%	×	25%	×	20%	=	5%
From Tier IV . . . to Tier III	63%	×	20%	×	80%	=	10%

[1] Tier IV data centers guarantee highest level of reliability: 99.995% uptime performance. Tiers III, II, and I deliver diminishing levels of performance, down to 99.671% reliability for Tier I.

5. How can we incorporate modular designs into our data center footprint?

There is a traditional model for data center expansion: enterprises build monolithic structures in a highly customized way to accommodate demand that is five or sometimes ten years out. In addition, they design facilities to meet the resiliency requirements of the most critical loads. New modular construction techniques (see sidebar “Three modular construction approaches”), however, have these advantages:

- shifting data center build programs from a craft process for custom-built capacity to an industrial process that allows companies to connect factory-built modules
- building capacity in much smaller increments
- making it easier to use lower-tier capacity
- avoiding the construction of new facilities, by leveraging existing investments (see sidebar “Deploying modular capacity: Two case studies”)

6. What is the complete list of key design decisions and their financial impact?

Even after the company has established its capacity requirements, dozens of design choices could substantially affect the cost to build. They include the following:

- redundancy level of electrical and mechanical equipment
- floor structure (for instance, single- or multistory)
- cooling technology (such as free-air cooling, evaporative chillers, and waterside economizers)
- degree to which components are shared between or dedicated to modules
- storm grade (for instance, the maximum wind speed a data center can withstand, as defined by regional or national standards, such as the Miami–Dade County building code and the Saffir–Simpson Hurricane Wind Scale)

Deploying modular capacity: Two case studies

A large government contractor was running out of critical capacity at a Tier III data center facility with an almost 40-year-old shell. This low-density space housed low-density computing platforms, as well as high-density virtual platforms, which ran at about 60 percent rack utilization. The company spent some $3 million to expand the mechanical and electrical plant to support an additional 500 kilowatts of critical load.

Those improvements helped the company to move high-density servers and storage equipment into in-space cooling units, which allowed it to run racks at 100 percent utilization while keeping the low-density equipment on an existing raised floor. This turned out to be much cheaper than building a brand-new 1-megawatt Tier III facility, which might have cost as much as $20 million. The company applied the same approach to its complete portfolio of data centers and avoided some $60 million to $80 million in capital costs as a result.

Modular facility-design techniques also help companies to segment applications into lower tiers within existing facilities. A large bank, for example, cut its build costs by up to 20 percent by using a modular, multitier facility design, which gives it the flexibility to change its facilities incrementally to meet the requirements of constantly evolving applications. As the software becomes more resilient, for example, the organization will be able to shift its data center portfolio to favor lower tiers.

Individual choices can have a disproportionate impact on costs per unit of capacity even after a company chooses its tier structure. A large global financial-services firm, for example, looked closely at its incident history and found that electrical failures—rather than mechanical ones—caused almost all of the issues at its facilities. Knowing this, the firm increased its electrical redundancy and decreased mechanical redundancy, shaving off several million dollars in construction costs per megawatt.

• • •

Given the scale of the investment required—billion-dollar data center programs are not unheard of—CIOs must undertake aggressive due diligence for their data center capital plans. They can often free up tens of millions of dollars from build programs by asking tough questions about resiliency, capacity, timing, tiering, and design. Perhaps more important, they can tell the executive committee and the board that they are using the company's capital in the most judicious way possible. ○

James Kaplan (James_Kaplan@McKinsey.com) is a principal in McKinsey's New York office, where **Allen Weinberg** (Allen_Weinberg@McKinsey.com) is a director; **Brent Smolinski** (Brent_Smolinski@McKinsey.com) is an associate principal in the Atlanta office.