Managing next-generation IT infrastructure

The days of building to order are over. The time is ripe for an industrial revolution.

James M. Kaplan, Markus Löffler, and Roger P. Roberts

In recent years, companies have worked hard to reduce the cost of the IT infrastructure—the data centers, networks, databases, and software tools that support businesses. These efforts to consolidate, standardize, and streamline assets, technologies, and processes have delivered major savings. Yet even the most effective cost-cutting program eventually hits a wall: the complexity of the infrastructure itself.

The root cause of this complexity is the build-to-order mind-set traditional in most IT organizations. The typical infrastructure may seem to be high tech but actually resembles an old-fashioned automobile: handmade by an expert craftsperson and customized to the specifications of an individual customer. Today an application developer typically specifies the exact server configuration for each application and the infrastructure group fulfills that request. The result: thousands of application silos, each with its own custom-configured hardware, and a jumble of often incompatible assets that greatly limit a company's flexibility and time to market. Since each server may be configured to meet an application’s peak demand, which is rarely attained, vast amounts of expensive capacity sit unused across the infrastructure at any given time. Moreover, applications are tightly linked to individual servers and storage devices, so the excess capacity can't be shared.

Now, however, technological advances—combined with new skills and management practices—allow companies to shed this build-to-order approach. A decade into the challenging transition to distributed computing, infrastructure groups are managing client-server and Web-centered architectures with growing authority. Companies are adopting standardized application platforms and development languages. And today's high-performance processors, storage units, and networks ensure that infrastructure elements rarely need hand-tuning to meet the requirements of applications.

In response to these changes, some leading companies are beginning to adopt an entirely new model of infrastructure management—more off-the-shelf than build-to-order. Instead of specifying the hardware and the configuration needed for a business application (“I need this particular maker, model, and configuration for my network-attached storage box . . .”), developers specify a service requirement (“I need storage with high-speed scalability . . .”); rather than building systems to order, infrastructure groups create portfolios of “productized,” reusable services. Streamlined, automated processes and technologies create a "factory" that delivers these products in optimal fashion (Exhibit 1). As product orders roll in, a factory manager monitors the infrastructure for capacity-planning and sourcing purposes.

With this model, filling an IT requirement is rather like shopping by catalog. A developer who needs a storage product, for instance, chooses from a portfolio of options, each described by service level (such as speed, capacity, or availability) and priced according to the infrastructure assets consumed (say, $7 a month for a gigabyte of managed storage). The system's transparency helps business users understand how demand drives the consumption and cost of resources.

Companies that make the transition gain big business benefits. By reducing...
complexity, eliminating redundant activity, and boosting the utilization of assets, they can make their infrastructure 20 to 30 percent more productive—on top of the benefit from previous efficiency efforts—thereby providing far greater output and flexibility. Even larger savings can be achieved by using low-cost, commodity assets when possible. Developers no longer must specify an application's technical underpinnings and can therefore focus on work that delivers greater business value; the new model improves times to market for new applications.

Nevertheless, making this transition calls for major organizational changes. Application developers must become adept at forecasting and managing demand so that, in turn, infrastructure groups can manage capacity more tightly. Infrastructure groups must develop new capabilities in product management and pricing as well as introduce new technologies such as grid computing and virtualization. As for CIOs, they must put in place a new model of governance to manage the new infrastructure organization.

The road forward

Deutsche Telekom knows firsthand the challenges involved: over 18 months, hoping to balance IT supply and demand, it implemented this new infrastructure-management model at two divisions (see sidebar, "Next-generation infrastructure at Deutsche Telekom"). In the old days, the company's IT infrastructure, like most, was a landscape of application silos. Today accurate forecasts of user demand are critical, so newly minted product managers must take a horizontal view, across applications, to assess the total needs of the business and create the right products. They must then work closely with infrastructure teams to align supply—infrastructure assets such as hardware, software, and storage—with demand.

In the past, employees of the infrastructure function were order takers. Now, they can be more entrepreneurial, choosing the mix of hardware, software, and technology that optimizes the infrastructure. To keep costs low, they can phase in grids of low-end servers, cheaper storage disks, and other commodity resources. Factory managers now focus on automating and "industrializing" production. Although Deutsche Telekom's two divisions didn't radically change their organizational or reporting structures, IT governance now seeks to ensure that product and service levels are consistent across business units in order to minimize costs and to improve the infrastructure's overall performance.
What we've seen at Deutsche Telekom and other companies suggests that creating a next-generation infrastructure involves action on three fronts: segmenting user demand, developing productlike services across business units, and creating shared factories to streamline the delivery of IT.

Segmenting user demand
Large IT organizations support thousands of applications, hundreds of physical sites, and tens of thousands of end users. All three of these elements are critical drivers of infrastructure demand: applications require servers and storage, sites need network connectivity, and users want access to desktops, laptops, PDAs, and so forth. To standardize these segments, an IT organization must first develop a deep understanding of the shape of current demand for infrastructure services and how that demand will most likely evolve. Then it needs to categorize demand into segments (such as uptime, throughput, and scalability) that are meaningful to business users.

When grouped in this way, most applications fall into a relatively small number of clusters. A pharmaceutical manufacturer, for instance, found that most of a business unit's existing and planned applications fell into one of five categories, including sales force applications that need around-the-clock support and off-line availability and enterprise applications that must scale up to thousands of users and handle batch transactions efficiently.

In contrast, a typical wholesale bank's application portfolio has more segments, with a wider range of needs. Some applications—such as derivatives, pricing, and risk-management tools—must execute computation-intensive analyses in minutes rather than hours. Funds-transfer applications allow for little or no downtime; program-trading applications must execute transactions in milliseconds or risk compromising trading strategies.

Although simple by comparison, the needs of physical sites and user groups can be categorized in a similar way. One marketing-services company that evaluated its network architecture, for example, segmented its sites into offices with more than 100 seats, those with 25 to 100, and remote branches with fewer than 25. A cable systems operator divided its users into senior executives with "concierge-support" needs, professional employees, call-center agents, and field technicians.

Most companies find that defining the specific infrastructure needs of applications, sites, and users is the key challenge of segmenting demand. Major issues include the time and frequency of need, the number of users, the amount of downtime that is acceptable, and the importance of speed, scalability, and mobility.

Standardizing products
Once the infrastructure group has assessed current and future demand, it can develop a set of productlike, reusable services for three segments: management and storage products for applications, access products such as desktops and laptops for end users, and network-access products for various sites. For each of these three product lines, the group must then make a series of decisions at both the portfolio and the product level.

At the portfolio level, it has to make decisions about the scope, depth, and breadth of product offerings, with an eye toward optimizing resources and minimizing costs. Exceptions must be detailed up front. The group may decide, for example, against offering products to support applications with stringent requirements, such as very-low-latency processing; these applications may be better built "by hand" and "from the ground up." Other applications, such as legacy ones, may be better left outside the new model if they're running well and can't easily be ported to new hardware. The group should also decide how to introduce new technologies and to migrate existing applications that are easier to move.

At the product level, the group must define the features, service levels, and price of each product. For each application support product, to give one example, it will be necessary to specify a programming language, an acceptable level of downtime, and a price for infrastructure usage. That price, in turn, depends on how the group decides to charge for computing, storage, processor, and network usage. The group has to consider whether its pricing model should offer discounts for accurate demand forecasts or drive users to specific products through strategic pricing.
Looking forward, companies may find that well-defined products and product portfolios are the single most important determinant of the infrastructure function's success. Developers and users may rebel if a portfolio offers too few choices, for instance, but a portfolio with too many won't reap the benefits of scale and reuse. Good initial research into user needs is critical, as it is for any consumer products company.

**The supply side: Creating shared factories**

The traditional build-to-order model limits the infrastructure function's ability to optimize service delivery. Delivery has three components: operational processes for deploying, running, and supporting applications and technologies; software tools for automating these operational processes; and facilities for housing people and assets.

At most companies, variations in architecture and technology make it impossible to use repeatable processes applied across systems. This problem hinders efficiency and automation and restricts the amount of work that can be performed remotely in low-cost locations, thus limiting the scope for additional cost savings.

In the next-generation infrastructure model, however, application developers specify a service need but have no input into the underlying technologies or processes chosen to meet it. The application may, for instance, require high-speed networked storage, but the developer neither knows nor cares which vendor provides the storage media. This concept isn't new—consumers who have call waiting on their home telephone lines don't know whether the local carrier has a Lucent Technology or Nortel Networks switch at its closest central office.

Because the infrastructure function can now choose which software technologies, hardware, and processes to use, it can rethink and redesign its delivery model for optimal efficiency. Using standardized and documented processes, it can start developing an integrated set of software tools to automate its operations. Next, by leveraging its processes and automation tools, it can develop an integrated location strategy that minimizes the need for data centers, so that more functions can operate remotely in low-cost—even offshore—locations.

**Building a new organization**

What changes must CIOs make to capitalize on these new opportunities? The next-generation infrastructure has major implications for the roles, responsibilities, and governance of the infrastructure organization.

The most critical new roles are those of the product manager, who defines products and product portfolios, and of the factory architect, who designs the shared processes to deploy, operate, and support them (Exhibit 2). Product managers must focus on service offerings and be accountable for reaching productivity targets. Their other key responsibilities include building relationships with business users and application developers, understanding and segmenting demand, defining product portfolios, and persuading developers and business users to accept their decisions.

Factory architects are, in equal parts, technology strategists and industrial engineers, codifying the architectures, processes, and tools that support the product portfolio. Their other key responsibilities include confirming that product commitments can be met, choosing technologies, defining processes, developing process-automation plans, and selecting tools. Although this was an established role at Deutsche Telekom, factory architects are now more focused on automating and industrializing production.

Organizational structures must change as well. Specialized silos with administrators focused on specific technology platforms—mainframes, midrange computing, distributed servers, storage, and voice and data networks—should give way to multidisciplinary teams that manage the performance of the infrastructure and the delivery of services.

CIOs must also put in place novel governance mechanisms to deal with capacity planning, the launch of new services, and investment-financing issues. Although Deutsche Telekom opted to keep its existing governance structure, many
companies create an enterprise-level infrastructure council to ensure the consistency of products and service levels across business units. Such consistency is critical for keeping costs low and optimizing performance.

To make sure the new infrastructure is running efficiently and to sustain performance improvements, IT leaders should focus on five key areas:

**1. Demand forecasting and capacity planning.** A key goal of the new infrastructure model is to match supply and demand more closely, thereby minimizing the waste of resources. To achieve this objective, the IT group must work closely with business units in order to forecast demand and thus improve capacity planning. Forecasts are more accurate when companies follow Deutsche Telekom’s example and aggregate demand across products instead of applications.

**2. Funding and budgeting.** Product demand drives budgets. Since the new model uses real demand forecasts, budgeting is easier. Moreover, with pricing transparency comes knowledge. Business units will now know what their IT choices are going to cost; the infrastructure group will understand the budget implications of user requests and be able to create a more accurate capital plan.

**3. Product-portfolio management.** Companies can expect to spend six months developing new-product portfolios. The infrastructure team should reexamine them two or three times during the first year to ensure that they are appropriate given projected workloads and emerging end-user needs. Thereafter, a yearly review usually suffices. Teams should monitor all phases of the product life cycle, from planning and sourcing new products to retiring old services and redeploying resources.

**4. Release management.** To ensure that new technologies or upgrades are integrated effectively and that change causes less upheaval and lost productivity, leading companies carefully manage the release of both infrastructure products and applications in parallel. Moreover, to plan ahead, application developers need to know about any impending change in the infrastructure catalog.

**5. Supply and vendor management.** IT leaders must ensure that computing resources are available to meet the contracted service levels of product portfolios. Infrastructure managers should revisit their sourcing strategy annually, seeking opportunities to lower costs and improve productivity.

Even with the restructuring and the new roles and processes in place, changing the build-to-order mind-set and culture may remain the biggest challenge of all. Deutsche Telekom adjusted its incentives, hired new people, developed training workshops, and appointed "change agents" to spread the word and build
enthusiasm. These organizational and cultural changes are central to realizing the potential of the next-generation infrastructure model. Investing the time and attention needed to get the right results is just as critical as refreshing the technical architecture.

Next-generation infrastructure at Deutsche Telekom

Otto Zeppenfeld’s cheerful demeanor may be surprising given his job. As head of IT operations for T-Com, Deutsche Telekom’s fixed-network division, he’s responsible for ensuring that all applications run smoothly, even during times of peak demand.

T-Com, which outsources almost all of its IT operations to its sister company T-Systems, provides voice and data services to about 40 million consumers and very small businesses. It generates higher revenues than any other division of Deutsche Telekom. The company’s IT infrastructure is massive: 1.3 petabytes of storage capacity, 25,000 MIPS of computing power, approximately 3,000 servers and 100,000 workstations, and hundreds of applications.

Many of T-Com’s IT infrastructure assets, like those of most companies, once sat idle waiting for peak loads. To address the problem, T-Com and T-Systems began implementing the key elements of a next-generation infrastructure model: productlike services, transparent pricing, strict demand forecasting, and capacity management.

Managing supply and demand across applications

For each infrastructure product category—storage, hosting, and so forth—T-Com appointed a product manager to assess demand across all applications in that category and to work with T-Systems on defining the right products and service levels and negotiating prices. T-Com’s process for forecasting demand aggregates it across all categories and then forwards that information to T-Systems for use in capacity planning and management.

T-Systems supplies T-Com’s products and manages the underlying hardware, software, and networks. Like T-Com, it takes a bird’s-eye view, looking across applications at total storage and computing needs. The two units now work in tandem to balance supply and demand—a radical departure from the traditional application-silo mentality.

The success of this model depends on two key factors: T-Com must learn to predict how much computing power it will need and when; T-Systems must learn how to use excess capacity in other areas. “T-Systems must take on a lot more responsibility,” notes Michael Auerbach, the T-Systems manager for all T-Com IT operations. “At the end of the day, it’s our job to leverage the idle capacity elsewhere.”

Paying only for usage

Since the new model requires T-Systems—rather than T-Com—to pick up the tab for unused capacity, T-Systems is under pressure to think and operate in new ways. Formerly, when Zeppenfeld needed a new business application for T-Com, T-Systems’ Auerbach supplied the appropriate hardware, software, and services and then tallied up the cost. It was usually the subject of intense debate because the value of complex computer systems is hard to determine, and so is the cost of the associated installation, operations, and maintenance.

Now, Zeppenfeld pays only for the computing resources he uses every month; the hardware, software, and storage needed to power T-Com’s applications are Auerbach’s problem. This model, however, gives T-Systems the freedom to make decisions that optimize the infrastructure as a whole rather than specific applications. Wherever possible, T-Systems uses cheaper commodity resources such as grids of low-end servers, storage disks, and Intel processors instead of Unix systems. In essence, it now acts more as an entrepreneur than as an order taker.

Transparency of costs is a major benefit of this model: T-Com merely reads off the bytes it consumed and pays a predetermined price that factors in T-Systems’ engineering support. T-Com’s invoice includes a handful of service categories (such as storage, backup, computing, the operation of applications, and the help desk) and quantifies usage in detail. Each service unit has a fixed
price, so T-Com knows exactly what it will pay for a gigabyte of storage, an hour of telephone support, or a backup copy of a database. Moreover, these services can be benchmarked individually, so TCom has the ability to check that prices are reasonable. Zeppenfeld and Auerbach agree that transparency helps create an atmosphere of trust.

**Gaining greater flexibility**

T-Com also gains flexibility: its contract lets it increase or decrease its computing capacity if it gives three months' notice. Drastic, across-the-board changes in usage are unlikely for most companies, but this added flexibility in individual areas is still a welcome benefit. Marketing, for instance, has fluctuating needs. Take an e-mail campaign generating several million responses. Previously, a six-month lead time was needed to purchase new hardware and software. The new model forces marketing to produce more accurate forecasts but cuts the lead time in half. Now, T-Com lets TSystems know about the marketing group's plans and requirements three months before such a campaign; the earlier it alerts T-Systems, the lower the added capacity costs. The department managers who, with their teams, plan in advance and make the most accurate forecasts can increase their savings. These incentives are designed to improve the forecasting of T-Com and the capacity planning of T-Systems.

**Making it happen**

A next-generation infrastructure model poses practical challenges. The concept of IT bills based on actual consumption, for instance, is still very much in the development phase. Moreover, some companies get stuck migrating legacy applications to the new systems. Depreciation schedules may mean that purchase and leasing agreements for old ones still have a long time to run. At most large companies, IT has hundreds of individual contracts expiring at different times, so it is hard to make a clean break.

T-Com and T-Systems found that an all-or-nothing approach was unnecessary. Making better use of existing resources and phasing in new technology allowed them to use savings generated by the new model to offset the cost of migrating applications. The two units also started small, focusing solely on a few key applications. Only later did they expand the model to computing services for mainframe applications. Today 80 percent of the relevant IT infrastructure has been converted.

T-Com and T-Systems are very satisfied with the early results of the new infrastructure model, which has delivered major cost savings, improved the use of assets, provided for greater flexibility, and made the system far less complex to manage.

**About the Authors**

James Kaplan is an associate principal in McKinsey’s global IT practice and specializes in IT infrastructure. He is based in New York. Markus Löffler is an associate principal in McKinsey’s global IT practice and specializes in IT infrastructure and architecture. He is based in Stuttgart. Roger Roberts leads McKinsey’s IT architecture practice in North America and specializes in the high-tech and industrial sectors. He is based in Silicon Valley.

The authors wish to thank Andrew Appel for his contributions to this article.

This article was first published in the Winter 2004 issue of *McKinsey on IT*.

**Notes**

1 Grid computing breaks down an application’s processing requirements into pieces for distribution among a number of servers. Server virtualization is a technology that allows a single central-processing unit to run a number of different operating systems—Windows NT, Windows XP, and Linux, for instance—at the same time.