Cloud and infrastructure
The end of the efficiency trade and the coming capacity crunch
perspective on what should be developed and controlled, and what needs to be monitored and governed.

To help equip the IT executive in forming those views and making those judgements we present points of view on key trends and topics.

Background

For the components in emerging technology to work, and for your IT portfolio to leverage them, compute hardware and cabling is still required. You can’t ignore what’s behind the façade of mobile devices and browser-based interfaces.

In this world of servers and data centers, IT executives have grown accustomed to the historical trend of being able to buy more compute horsepower each year for the same or less money, in smaller and smaller sizes. This trend has enabled efficiency trades between the space and power required to deliver increasing compute and storage to the business.

Given that compute and IT facilities costs still account for much of the IT budget (either yours or your service provider’s), changes in these trends may create strategic capacity problems and drive up long-term costs.

In this article we discuss ways compute refreshes will likely start to lose value, that technology advances may soon counter standardization moves, and that data center consolidation will likely need to reverse.

We will discuss some possible mitigations, and look closely at developments in that largest cost component—data centers.

About the series

Overview

In order to enable new business and preserve existing value, global Information Technology (IT) executives should address the growing dichotomy between the agility companies want, and the stability they need. This dichotomy is becoming exacerbated as legacy data centers become farther and farther removed from the cloud and from mobile end users. Meanwhile, delivering data to devices and taking advantage of platforms often means giving up some control over operating systems, hardware, and data centers. It’s 9 a.m. on Monday. Do you know where your data is?

Introduction

As IT executives look to provide value from their IT portfolios, they are balancing a mix of emerging, current, and legacy technologies. With the world of technology ever evolving at pace, the gulf between emerging and legacy continues to widen. The consumer market drives advances in end user devices and end user expectations. Many service vendors invest in cloud, virtualization, and orchestration, while manufacturers attempt to deliver more compute horsepower at lower cost. Behind the tablets, clouds, and chips, there still sits a data center and a room full of legacy infrastructure waiting for refresh.

This widening gap between end user devices, data mobility, cloud services, and back office legacy systems challenges the IT executive to manage and maintain technology in a complex array of delivery capabilities. From mobile apps to mainframe MIPS, and from in-house servers to sourced vendor services, managing this broad range requires a view on how much can change by when, an appropriate operating model, and a balanced perspective on what should be developed and controlled, and what needs to be monitored and governed.

To help equip the IT executive in forming those views and making those judgements we present points of view on key trends and topics.

Background

For the components in emerging technology to work, and for your IT portfolio to leverage them, compute hardware and cabling is still required. You can’t ignore what’s behind the façade of mobile devices and browser-based interfaces.

In this world of servers and data centers, IT executives have grown accustomed to the historical trend of being able to buy more compute horsepower each year for the same or less money, in smaller and smaller sizes. This trend has enabled efficiency trades between the space and power required to deliver increasing compute and storage to the business.

Given that compute and IT facilities costs still account for much of the IT budget (either yours or your service provider’s), changes in these trends may create strategic capacity problems and drive up long-term costs.

In this article we discuss ways compute refreshes will likely start to lose value, that technology advances may soon counter standardization moves, and that data center consolidation will likely need to reverse. We will discuss some possible mitigations, and look closely at developments in that largest cost component—data centers.
We loved the efficiency trade

Since the 1960s, technical advances provided computing capability at lower cost and lower electrical usage. The economics of what we commonly refer to as “Moore’s Law” have been driven by a variety of factors such as chip manufacturers getting more transistors into the same space (classic Moore’s), the power density of transistors scaling downward with transistor size (Dennard Scaling), and increases in the electrical efficiency of computation (Koomey’s Law). For IT executives these trends presented some interesting opportunities. By refreshing hardware in cycles:

- An application could be serviced by IT infrastructure with a smaller form factor inside the data center.
- The same business functionality could be presented with better compute performance for less operating cost.
- By reducing the physical and electrical footprint of an application inside the data center, one could choose to consolidate space, or grow business functions in the existing available capacity.

In practical terms, this often meant that by reserving some CapEx every year for new hardware, the OpEx budgets for IT could reduce or remain flat, while delivering the same or more business functionality. The business has become used to getting more for less.

This trade is, however, starting to run its course as both advances in transistor technology and compute equipment pricing slow down.

Compute advances are slowing down

Moore’s Law is breaking down as transistor sizes are approaching their practical physical limitations in silicon. Dennard Scaling has actually broken down and heat dissipation at smaller sizes and higher frequencies is now a problem. A corollary phenomenon is that manufacturing costs increase as the technology becomes more difficult to produce which results in slower delivery cycles of new chips. We can see the evidence of this at work in terms of the transition from 22 to 14 nanometer semiconductor production.

Prices are being affected

These changes are now also having a marked economic impact. The long term historical trend of the quality adjusted price of IT equipment is that it continually decreases, following a similar curve to Moore’s Law. But the price trend has seen a remarkable slowdown over the past four years, as evidenced by Federal Reserve figures.
Manufacturers are responding

In recent years, manufacturers have delivered performance and cost improvements with features such as multi-core architectures and specialized chips over increased transistor density. And given the upfront investments and manufacturing costs, it is becoming more advantageous for server and device manufacturers to work with existing chip platforms longer.

As the 50 year cycle of silicon chip advances begins to slow down and reach its physical limits, there is active research and development in subsequent and replacement technologies. For example:

- Silicon may be replaced by other elements—perhaps graphene or carbon nanotubes.
- Work on “memristors” continues with the promise of breaking down the memory wall with essentially a non-volatile RAM.
- Replacing the electron with the photon for data transfer is the notion behind silicon photonics.
- Moving from the bit to the “qubit” could leverage quantum phenomena for faster problem solving.

But refresh starts to lose value

Regardless of these innovations, there is little promise that these or other step-change technologies are yet close to mass mainstream manufacture at competitive prices.

So compute power advances are slowing, price improvements are stalling, and replacement technologies not yet ready. This results in a slower refresh cycle and a longer wait for the usual efficiency gain.

Multicore architectures

If there is an increasing reliance on multiple cores, then your software development should align to techniques and languages best suited to leverage multicore and parallel processing. This will not necessarily suit all use cases though, as you likely have business needs and applications where independent reading is not advantageous. Additionally, it can add complexity in terms of thread safety, concurrence, memory constraints, and parallel.

Performance through specialization

To continue to deliver compute performance gains, manufacturers can respond with the introduction of more specialized or single purpose chips or hardware components.

This could mean that the trend to commoditize infrastructure could also break down. Manufacturers should consider trying new approaches and new technologies to increase performance and may need to move away from the standard chipset architectures and processor, memory and data relationships.

Many IT executives have worked to converge on commodity infrastructures and streamline and optimize support and operations by not having too many components and versions requiring specific support regimes and skills. But if manufacturers bring to market more specialized chips and components, then keeping current means adding new and specialized technologies into IT portfolios on top of all the existing and legacy items.

This potential to fuel a move to add diversity in the tech landscape is a counter to many current strategies. And it means adding development, support, and maintenance capabilities to help manage the specific new items, and the overall added complexity of the landscape.

The result is however, another driver for increasing costs and overhead. If there is any silver lining to this perspective, it could be the potential to use assets for longer. If performance improvements are not so quick to materialize as in the past, and increasing complexity of new special components is not presenting sufficient value, then using what you have longer is a strategy. Extending infrastructure and application refresh lifecycles may also provide opportunities to shift some priorities, for example onto aging legacy refresh.

Multicore architectures

If there is an increasing reliance on multiple cores, then your software development should align to techniques and languages best suited to leverage multicore and parallel processing. This will not necessarily suit all use cases though, as you likely have business needs and applications where independent reading is not advantageous. Additionally, it can add complexity in terms of thread safety, concurrence, memory constraints, and parallel.

Performance through specialization

To continue to deliver compute performance gains, manufacturers can respond with the introduction of more specialized or single purpose chips or hardware components.

This could mean that the trend to commoditize infrastructure could also break down. Manufacturers should consider trying new approaches and new technologies to increase performance and may need to move away from the standard chipset architectures and processor, memory and data relationships.

Many IT executives have worked to converge on commodity infrastructures and streamline and optimize support and operations by not having too many components and versions requiring specific support regimes and skills. But if manufacturers bring to market more specialized chips and components, then keeping current means adding new and specialized technologies into IT portfolios on top of all the existing and legacy items.

This potential to fuel a move to add diversity in the tech landscape is a counter to many current strategies. And it means adding development, support, and maintenance capabilities to help manage the specific new items, and the overall added complexity of the landscape.

The result is however, another driver for increasing costs and overhead. If there is any silver lining to this perspective, it could be the potential to use assets for longer. If performance improvements are not so quick to materialize as in the past, and increasing complexity of new special components is not presenting sufficient value, then using what you have longer is a strategy. Extending infrastructure and application refresh lifecycles may also provide opportunities to shift some priorities, for example onto aging legacy refresh.
The coming capacity crunch

In the 1990s and early 2000s IT managers were faced with huge increases in demand for IT compute capability due to the explosion of the internet and World Wide Web usage. This created new industries and delivery channels that enabled more data to be consumed and processed. Growing compute requirements corresponded to an increase in corporate data center capacity.

The initial investment in the nascent digital economy followed traditional cost trends. Compute size reduction, power efficiency, performance gains, and then virtualization, together enabled capacity needs to decrease faster than net new demand inside the organization. The result was an opportunity to neutralize physical space growth, consolidate compute, storage, and rack space, and create surplus data center capacity.

Some have responded in recent years with corporate programs aimed at reducing and consolidating their data centers—betting that continued compute and storage advancements would net off against new business.

But the slowdown of microprocessor performance gains and size reductions means that once again the capacity curve may invert. If you can’t easily buy more compute power in the same server footprint anymore, then increasing compute power can only mean more servers of the same size, more space needed, and more electricity consumed.

Optimization programs and data center consolidation plays may soon run their course. Refreshing equipment may not get you more compute for less, and the increasing needs of your business will likely require an increase in rack unit consumption.

There may be a coming capacity crunch in the data center. Do you respond by building new capacity or leveraging someone else’s?

The capacity response: build or buy?

With the capacity crunch on the horizon, the CIO should be considering these dynamics:

- Economies of scale favor large scale DC builds. IDC predicts that by 2017, there will be only 6 to 8 mega global players in IaaS, maybe fewer. 10
- Data center construction techniques and power sources evolve.
- Yet design lifecycles of data center buildings outlast compute and storage technology improvements by an order of magnitude.
- Following shorter term technology trends can result in misaligned long term capital investments.
- On the other hand, real estate ownership, balance sheet assets, control of capacity, and securing the risk profile internally, have advantages for some business models.

- The longer term danger is that when new technologies do become mainstream cost effective, there likely will be again an opportunity to right size capacity to that new dynamic (e.g., x86 virtualization).

Whether build or buy, close consideration of all service, cost, and risk parameters are a central component of your data center strategy. We examine some key DC trends that affect build and buy investments and running costs.

CoLo or not to CoLo

One of the largest costs you or your service provider may have is wrapped up in IT facilities. Building and maintaining enterprise data centers is usually an industrial scale undertaking at industrial scale cost. Here, economies of scale and lengthy lifecycles are key to managing unit costs.

Data center construction continues to grow at a pace largely driven by the primary wholesale colocation market11. New builds tend to be at a large scale for data center colocation providers or large scale internet based service providers such as Facebook, Microsoft, and Amazon. CoLo providers are investing in site selection, power and network capabilities, a minimum of Tier III resilience, and substantial square footage. It is sensible to have a hard look at the capital and time required to build your own, and weigh that against a burgeoning market of substantial scale.

One method many CoLo and webscale service providers now use is DC modularization. This growing trend is about delivering data center capacity in smaller units. Pre-fabricated data center modules can be ordered and delivered as needed—better aligning spend and capacity needs, and reducing the massive up front capital investments into large scale bricks and mortar12.

One of the questions to ask as an IT leader is: Can I compete on cost and risk for my 25-50k sq. ft. needs and 2-5 MW of IT power with respect to providers building out many multiples of those numbers? As seen in the DC snapshot above, one service provider data center can be larger in space and power than your entire global estate.

---

**Individual DC sizes: Space snapshot**

- IC: 830,000 sq. ft.
- DRT: 1.1M sq. ft.
- Facebook: 600,000 sq. ft
- Microsoft: 500,000 sq. ft.
- Equinix: 226,000 sq. ft. 13

---

One of the largest costs you or your service provider may have is wrapped up in IT facilities. Building and maintaining enterprise data centers is usually an industrial scale undertaking at industrial scale cost. Here, economies of scale and lengthy lifecycles are key to managing unit costs.

Data center construction continues to grow at a pace largely driven by the primary wholesale colocation market11. New builds tend to be at a large scale for data center colocation providers or large scale internet based service providers such as Facebook, Microsoft, and Amazon. CoLo providers are investing in site selection, power and network capabilities, a minimum of Tier III resilience, and substantial square footage. It is sensible to have a hard look at the capital and time required to build your own, and weigh that against a burgeoning market of substantial scale.

One method many CoLo and webscale service providers now use is DC modularization. This growing trend is about delivering data center capacity in smaller units. Pre-fabricated data center modules can be ordered and delivered as needed—better aligning spend and capacity needs, and reducing the massive up front capital investments into large scale bricks and mortar12.

One of the questions to ask as an IT leader is: Can I compete on cost and risk for my 25-50k sq. ft. needs and 2-5 MW of IT power with respect to providers building out many multiples of those numbers? As seen in the DC snapshot above, one service provider data center can be larger in space and power than your entire global estate.
Power efficiency

Optimizing the electricity bill
Looking at the cost components of data centers, your own, or a service provider, a major operating cost is electricity. In a traditional raised floor data center for every dollar spent on powering a server, another dollar is spent on the cooling infrastructure to keep the server at safe operating temperatures. This $1 + $1 equals a power usage effectiveness (PUE) of 2. Every cent of that cooling dollar is wasted overhead cost not adding any business value. How can you or the service provider reduce that cooling dollar as much as possible? Strive for a PUE of only 1 or less — where the only dollar spent is the electrical power for the server itself, or in a PUE less than 1 where residual heat energy is put to good use.

New cooling technology
There are a variety of ways to optimize heat and cooling flow in a traditional site that reduce that wasted dollar to 70 cents or even 50 cents — driving PUE to 1.5 or so. Getting lower than that takes some new approaches. Most commonly applied now is designing to avoid a cooling infrastructure entirely, and using only free air-flow. Intelligent use of air pressure differentials and flow across the compute racks can largely eliminate the need for air-conditioning and chilled water units. The most effective of all is in smaller purpose built data center modules that are hyper efficient in their management of heat and air. These technologies can push that wasted dollar down to 15 or even 10 cents—a PUE of 1.1.

Off-grid power sources
Another trend to consider is power generation. Primary power to the data center likely comes off the main grid—ideally via multiple paths, or even from multiple grids. Backup power for resilience may come from a combination of batteries, generators, and flywheels. Aligned with corporate social responsibility, often the question now is — how is that power generated? Primary power delivered from energy providers can be varying shades of green, covering the range from nuclear, hydro, oil, to coal, with some wind and solar sources thrown in. Secondary sources may be diesel, but also increasingly solar, wind, fuel-cell, new battery technologies, and other emerging methods.

There is the consideration some are having as well in using direct current for the DC, as opposed to alternating current. Large scale service providers, in addition to building their own pre-fab data centers, are looking to become off-grid energy independent.

All of these data center development approaches influence the quality, resilience, and the cost of either your own team, or of a service provider, which indirectly influences the price charged by your cloud vendor.

Conclusion
In this paper we have examined ways slowdowns in the development of new transistor technology could drive a number of unintended consequences:
• Refresh cycles could slow
• Capacity needs might increase
• Complexities would likely return
• Price could rise

If these consequences materialize, the knock on effects to your compute, support, and data center strategies need to be thought through. The next article in our series considers the operating aspects of this and our other trends — how does the IT executive consider the optimal operating model for managing emergent and legacy technologies?

For more information
Ranjit Bawa
Principal
Deloitte Consulting LLP
rbawa@deloitte.com

Myke Miller
Managing Director
Deloitte Consulting LLP
mykemiller@deloitte.com

Shawn Lund
Managing Director
Deloitte Consulting LLP
shlund@deloitte.com

Contributor
Richard Pone
rpane@deloitte.com
This publication contains general information only and is based on the experiences and research of Deloitte practitioners. Deloitte is not, by means of this publication, rendering business, financial, investment, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte, its affiliates, and related entities shall not be responsible for any loss sustained by any person who relies on this publication.

About Deloitte
Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee, and its network of member firms, each of which is a legally separate and independent entity. Please see www.deloitte.com/about for a detailed description of the legal structure of Deloitte Touche Tohmatsu Limited and its member firms. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

© 2016 Deloitte Development LLC. All rights reserved.