# Where the cloud meets the ground

## Data centres are quickly evolving into service factories

It IS almost as easy as plugging in a laser printer. Up to 2,500 servers—in essence, souped-up personal computers- are crammed into a 40-foot {13 metre) shipping container. A truck places the container inside a bare steel-and-concrete building. Workers quickly connect it to the electric grid, the computer network and a water supply for cooling. The necessary software is downloaded automatically. Within four days all the servers are ready to dish up videos, send e-mails or crunch a firm's customer data.

This is Microsoft's new data centre in Northlake, a suburb of Chicago, one of the world's most modern, biggest and most expensive, covering 500,000 square feet (46,000 square metres) and costing $5oom. One day it will hold 400,000 servers. The entire first floor will be filled with 200 containers like this one. Michael Manos, the head of Microsoft's data centres, is really excited about these containers. They solve many of the problems that tend to crop up when putting up huge data centres: how to package and transport servers cheaply, how to limit their appetite for energy and how to install them only when they are needed to avoid leaving expensive assets idle.

But containers are not the only innovation of which Mr Manos is proud. Microsoft's data centres in Chicago and across the world are equipped with software that tells him exactly how much power each application consumes and how much carbonit emits. "We're building a global information utility," he says.

Engineers must have spoken with similar passion when the first moving assembly lines were installed in car factories almost a century ago, and Microsoft's data centre in Northlake, just like Henry Ford's first large factory in Highland Park, Michigan, may one day be seen as a symbol of a new industrial era.

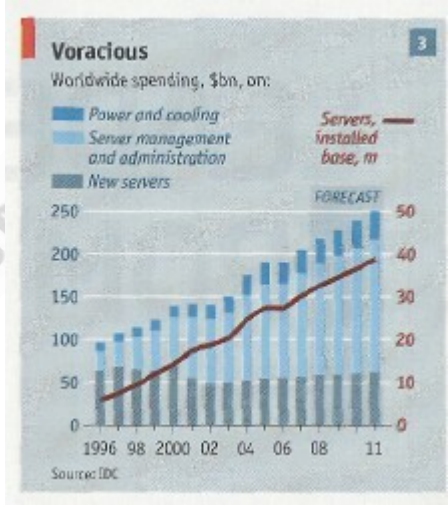
Before Ford revolutionised carmaking, automobiles were put together by teams of highly skilled craftsmen in custom-built workshops. Similarly, most corporate data centres today house armies of "systems administrators", the craftsmen of the information age. There are an estimated 7,000 such data centres in America alone, most of them one-off designs that have grown over the years, reflecting the history of both technology and the particular use to which it is being put. It is no surprise that they are egregiously inefficient. On average only 6% of server capacity is used, according to a study by McKinsey, a consultancy, and the Uptime Institute, a think-tank. Nearly 30% are no longer in use at all, but no one has bothered to remove them. Often nobody knows which application is running on which server. A widely used method to find out is: "Let's pull the plug and see who calls."

Limited technology and misplaced incentives are to blame. Windows, the most pervasive operating system used in data centres, allows only one application to run on any one server because otherwise it might crash. So IT departments just kept adding machines when new applications were needed, leading to a condition known as "server sprawl" (see chart 3). This made sense at the time: servers were cheap, and ever-rising electricity bills were generally charged to a company's facilities budget rather than to IT.

To understand the technology needed to industrialise data centres, it helps to look at the history of electricity. It was only after the widespread deployment of the "rotary converter", a device that transforms one kind of current into another, that different power plants and generators could be assembled into a universal grid. Similarly, a technology called "virtualisation" now allows physically separate computer systems to act as one.

### Virtually new

The origins of virtualisation go back to the 19605, when IBM developed the technology so that its customers couldmake better use of their mainframes. Yet it lingered in obscurity until VMware, now one of the world's biggest software firms, applied it to the commodity computers in today's data centres. It did that by developing a small program called hypervisor, a sort of electronic traffic cop that controls access to a computer's processor and memory. It allows servers to be split into several "virtual machines", each of which can run its own operating system and application.




"In a way, we're cleaning up Microsoft's sins," says Paul Maritz, VMware's boss and a Microsoft veteran, "and in doing so we're separating the computing workload from the hardware." Once computers have become more or less disembodied, all sorts of possibilities open up. Virtual machines can be fired up in minutes. They can be moved around while running, perhaps to concentrate them on one server to save energy. They can have an identical twin which takes over should the original fail. And they can be sokl'prepackaged as "virtual appliances".

viwware and its competitors, which now include Microsoft, hope eventually to turn a data centre-or even several of them-into a single pool of computing, storage and networking resources that can be allocated as needed. Such a "real-time infrastructure", as Thomas Bittman of Gartner calls it, is still years off. But the necessary software is starting to become available. In September, for instance, VMware launched a new "virtual data-centre operating system".

Perhaps surprisingly, it is Amazon, a big online retailer, that shows where things are heading. In 2006 it started offering a computing utility called Amazon Web Services (AWS). Anybody with a credit card can start, say, a virtual machine on Amazon's vast computer system to run an application, such as a web-based service. Developers can quickly add extra machines

when needed and shut them down if there is no demand (which is why the utility is called Elastic Computing Cloud, or EC2). And the service is cheap: a virtual machine, for instance, starts at 10 cents per hour.

If Amazon has become a cloud-computing pioneer, it is because it sees itself as a technology company. As it branched out into more and more retail categories, it had to develop a sophisticated computing platform which it is now offering as a service for a fee. "Of course this has nothing to do with selling books," says Adam Selipsky, in charge of product management at AWS, "but it has a lot to do with the same technology we are using to sell books."

Yet Amazon is not the only big online company to offer the use of industrial-scale data centres. Google is said to be operating a global network of about three dozen data centres loaded with more than 2m servers (although it will not confirm this). Microsoft is investing billions and adding up to 35,000 servers a month. Other internet giants, such as Yahoo!, are also busy building huge server farms.

In some places this has led to a veritable data-centre construction boom. Half a dozen are being built in Quincy, a hamlet in the middle of America's Washington state, close to the Columbia River. The attraction is that its dams produce plenty of low-cost power, which apart from IT gear is the main input for these computing farms. On average, cooling takes as much power as computing. Microsoft's new data centre near Chicago, for instance, has three substations with a total capacity of 19SMW, as much as a small aluminium smelter.

But cheap electricity is only one, albeit important, criterion for choosing the site of a data centre. Microsoft currently feeds 35 sets of data into an electronic map of the world, including internet connectivity, the availability of IT workers, even the air quality (dry air makes a good coolant), to see where conditions are favourable and which places should be avoided. Apparently Siberia comes out well.

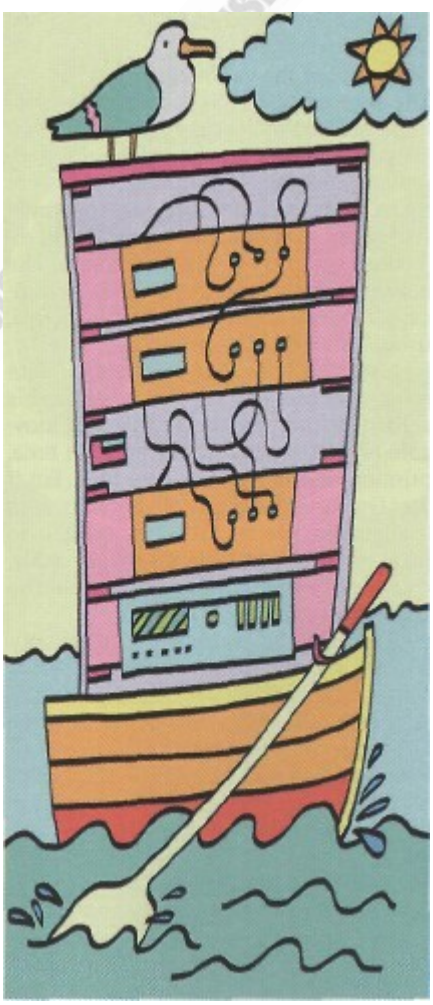
Google, for its part, seems to be thinking of moving offshore. In August it applied for a patent for water-based data centres. "Computing centres are located on a ship or ships, anchored in a water body from which energy from natural motion of the water may be captured, and turned into electricity and/or pumping power for cooling pumps to carry heat away,[1]" says the patent application.

Many chief information officers would love to take their IT infrastructure out to sea and perhaps drown it there. Even as demand for corporate computing continues to increase, IT budgets are being cut. At the same time many firms' existing IT infrastructure is bursting at the seams. According to IDC, a market-research firm, a quarter of corporate data centres in America have run out of space for more servers. For others cooling has become a big constraint. And often utilities cannot provide the extra power needed for an expansion.

**Fewer,** bigger, better

So IDC thinks that many data centres will be consolidated and given a big makeover. The industry itself is taking the lead. For example, Hewlett-Packard (HP) used to have 85 data centres with 19,000 IT workers worldwide, but expects to cut this down to six facilities in America with just 8,000 employees by the end of this year, reducing its IT budget from 4% to 2% of revenue.

Other large organisations are following suit. Using VMware's software, BT, a telecoms firm, has cut the number of servers in its 57 data centres across the world from 16,000 to 10,000 yet increased their workload. The us Marine Corps is reducing the number of its IT sites from 175 to about 100. Both organisations are also starting to build internal clouds so they can move applications around. Ever more firms are expected to start building similar in-house, or "private", clouds. The current economic malaise may speed up this trend as companies strive to become more efficient.



But to what extent will companies outsource their computing to "public" clouds, such as Amazon's? James Staten of Forrester Research, a market-research firm, says the economics are compelling, particularly for smaller firms. Cloud providers, he says, have more expertise in running data centres and benefit from a larger infrastructure. Yet many firms will not let company data float around in a public cloud where they could end up in the wrong hands. The conclusion of this report will consider the question of security in more detail.

It does not help that Amazon and Google recently made headlines with service interruptions. Few cloud providers today offer any assurances on things like continuity of service or security (called "service-level agreements", or SLAS) or take on liability to back them up.

As a result, says Mr Staten, cloud computing has not yet moved much beyond the early-adopter phase, meaning that only a few of the bigger companies are using it, and then only for projects that do not critically affect their business. The Washington Post, for instance, used Amazon's AWS to turn Hillary Clinton's White House schedule during her husband's time in office, with more than 17,000 pages, into a searchable database within 24 hours. NASDAQ uses it to power its service providing historical stockmarket information, called Market Replay.

Stefan van Overtveldt, the man in charge of transforming BT'S IT infrastructure, thinks that to attract more customers, service providers will have to offer "virtual private clouds", fenced off within a public cloud. BT plans to offer these as a service for firms that quickly need extra capacity.

So there will be not just one cloud but a number of different sorts: private ones and public ones, which themselves will divide into general-purpose and specialised ones. Cisco, a leading maker of networking gear, is already talking of an "intercloud", a federation of all kinds of clouds, in the same way that the internet is a network of networks. And all of those clouds will be full of applications and services. •