High-availability Internet Data Centre Concepts

by Gregory J. Graham
Telecom Marketing Associates, Kanata, ON

Abstract

As services based on the Internet have become part of everyday life for Canadians, the infrastructure that provides these services must scale upwards in robustness and capacity. Lessons learnt from the public switched telephone network are being applied, with new twists, to Internet infrastructure. The importance of providing secure, high-availability infrastructure has become even more obvious following the recent events of September 11, 2001.

This article describes leading approaches to designing high-availability data centres used for web-hosting, e-commerce, storage and other IP applications.

Sommaire

Comme les services basés sur l'Internet font maintenant partie de la vie quotidienne des canadiens, l'infrastructure qui assure ces services doit être “scale upwards” en robustesse et en capacité. Les leçons apprises des réseaux téléphoniques publics à interrupteurs sont appliqués à l'infrastructure de l'Internet avec des variantes. Suite aux évènements du 11 septembre 2001, il est devenu encore plus important d'offrir une infrastructure sécuritaire et a disponibilité élevée. Cet article décrit les approches d'avant-garde dans la conception de centres de données à haute disponibilité utilisés pour hôtes de pages web, commerce électronique, entreposage de données et autres applications utilisant le protocole Internet.

1.0 Introduction

With today's growing Internet use, 24x7 schedules and global availability via the web, mission-critical corporate web sites and e-commerce operations simply can't stop working. Down time clearly costs money, either through lost customers, revenue, or productivity.

Internet Data Centres (IDCs) are carefully-designed installations which combine central office reliability and data processing centre concepts with state-of-the-art physical security. The result? An extremely robust infrastructure. In fact, leading IDCs are meticulously designed to do just two things:

- Offer always-up service by minimizing the risk of any outages, and
- If an outage were ever to occur, recover as quickly as possible.

However, this high-availability infrastructure comes at a high cost. Internet data centres require carefully-selected real estate, a custom-purpose building, physical and IT security systems, environmental control systems, standby power capability, and complete redundancy. Sharing these costs between many users decreases the per-user cost to an affordable level, which is the fundamental concept behind the Internet data centre business model.

This article looks at the concepts behind Internet data centres, and how risk is designed out of an IDC right from the beginning.

2.0 What is Inside a Data Centre?

What's inside a data centre depends on its business model. In general, IDCs make their money from:

- Renting space, power and connectivity,
- Operating outsourced IT operations, and
- Offering a suite of optional professional services.

2.1 Carrier Hotels

A data centre offer can be as simple as providing space to communication service providers who install their networking equipment in it. The service provider rents temperature-controlled space with power and connectivity.

Location is key since proximity to the service provider's customers, and redundant fibre connections, is desirable. This is especially important when data centres are located in crowded downtown cores where availability to these items can be limited.

Note that the service providers are responsible for the installation and operation of their equipment and service. The data centre operator is merely a landlord.

When more than one communications carrier offers connection to the same data centre, it becomes a “carrier hotel”. This interconnectivity is used by the carriers themselves to exchange off-net traffic. This benefits their customers, too. In a carrier hotel, customers have a greater choice of potential suppliers -- they can connect to any carrier in the carrier hotel without relocating their equipment.

2.2 Co-Location

The business practice of deploying and operating equipment on someone else's premises is called “co-location”. A data centre operated by a carrier can generate co-location revenue by providing space and connectivity to competitive local exchange carriers, competitive availability provides, and Internet service providers. These tenants obtain connections to the public switched telephone network without the expense of developing their own costly facilities.

Devices such as switches, gateways and availability multiplexors reside in the central office and connect directly to the local loop. Redundant trunk lines guarantee high availability under almost any conditions. Traditional central offices (CO) were thus the first data centres. The primary difference is that IDCs house servers while COs house switches and availability equipment.

2.3 Managed Hosting

Unlike co-location, managed hosting is a turn-key service. Customers outsource their web business to the operator of the Internet data centre. In return, IDC operators provide their customers with quality guarantees via Service Level Agreements (SLAs) -- contracts specifying pricing, terms, metrics, and performance penalties. Equipment can be owned either by the customer or the IDC operator, depending upon the commercial arrangements.

A twist on managed hosting is dedicated hosting, a operational approach that dedicates distinct servers for each customer. This eliminates potential issues such as software conflicts and the difficulty of scheduling maintenance for shared equipment.

2.4 Value-Added Services

Internet data centres have supplemented these basic business models with a host of à la carte services. Professional services offered often include firewalls, network monitoring, load balancing, performance testing, security audits, private data networks and database services. These services all require networking expertise which the customer may not have in-house.

3.0 From The Central Office To The IDC

Central offices (COs) have long been outfitted with heating, air conditioning, massive battery backup systems, and redundant trunk lines to supply conditioned space, power, security, and network connectivity.

Because central offices terminate telephone local loops on switches, switch ports are the basic unit of CO operations. A typical central office in North America might terminate 10,000 local loops, and thus require a switch with 10,000 ports.

However, enterprises need physical facilities to house the servers that deliver their web content, e-commerce, and customer relationship management systems. These systems more frequently operate on 110 VAC rather than the 48 VDC power traditionally available in a central office. This is a key difference between a CO and an IDC.

Space within an IDC can be separated into areas containing equipment racks, electrical power, and connectivity. Depending on the customer's wishes and management philosophy, equipment racks may be standalone, ganged in rows, in locked cages, or even in separate locked vaults.

Regardless of how space is rented, customers' equipment usually ends up mounted in a rack. A rack is the basic unit of data centre operations since space utilization -- rather than switch ports -- drives revenues and costs. The footprint of a rack ranges from 20 to 30 square feet (1.85 to 2.79 square metres); reducing this footprint increases the density of racks and thus potential revenues.

Density can also be increased by extending rack height. Server manufacturers have continually reduced the physical size of servers. While the latest servers resemble a pizza box and are as little as 1.75 inches high (4.45 cm or 1 rack unit; a standard 78-inch high rack has 45 rack units), servers vary widely in their size and processing power. Very large servers are supplied with stand-alone custom cabinets while over forty of the latest pizza-box server appliances can fit into a single rack.

This goal of fitting a maximal number of servers into a minimal volume of space creates severe electrical supply and cooling problems which did not exist for central offices. COs of the past had limited computing power and large numbers of space-consuming I/O ports. In contrast, data centres have staggering amounts of computing power which dissipates enormous amounts of heat.

4.0 Designing Out Risk Factors

4.1 Location

Designing out risk starts with the location of a data centre. Selection of a site can minimize the risk of natural disaster by avoiding flood zones and areas with frequent tornadoes, etc. Data centres should also be located a safe distance from highways, railroad tracks, and flight paths.

Availability of optical fibre communications facilities is another consideration. Ideally, an IDC should be situated so that it can connect to redundant fibre connections from more than one carrier.

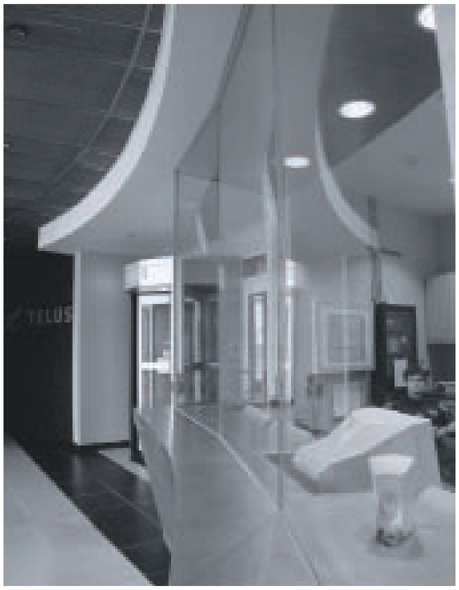

Figure 1 (left): The hardened lobby of a TELUS intelligent Internet Data Centre has bullet-resistant glass and walls, a locked mantrap, biometric scanners, and 24x7 security guards.

4.2 Construction

Next, the building envelope of a data centre should be constructed to withstand both natural disasters and attack. Although such design practises should be routine, these features are increasingly sought in today's post 9/11 reality. Internet data centers usually have not just thick concrete walls, a minimum of windows, bullet-resistant glazing, bullet-resistant lobby walls with steel-plating concealed under the drywall, and a locking mantrap entrance, but may even be structurally-reinforced to withstand bomb blasts and earthquakes (Figure 1).

As noted above, increasing the amount of equipment in each rack improves the economics of a data centre. But increasing the rack or server density may require thicker floor slabs. The soil under the building must also be able to carry exceptionally heavy loads, so new construction must be specified accordingly. An existing building being considered for retrofit should be carefully inspected by structural experts to ensure that it can meet any increased load requirements.

4.3 Physical Security

High levels of both physical and IT security are fundamental for IDCs. Typical physical security measures include perimeter fences, motion sensors, alarms, guards, video surveillance, and controlled availability to equipment.

In particular, security staff and measures must prevent customers from having physical availability to each others’ equipment, especially competitors’ equipment.

For example, in TELUS data centres, persons attempting to gain entrance must pass several levels of scrutiny. First, customers provide a list of personnel for whom they wish to authorize physical access to their equipment in the data centre. All customer personnel must pass a criminal background check by local police authorities. Next, authorized personnel are issued a photo ID card with a swipestrip. Finally, biometric data in the form of a fingerprint is collected.

Someone entering the data centre must thus appear on the list, present the photo ID card to security guards, have it verified by swiping it in a card reader, scan their fingerprint, and only then can pass through a mantrap which is normally kept locked (Figures 2 and 3).

Once inside, a visitor is not left free to roam throughout the data centre. Physical availability to unauthorized areas of the IDC is prevented since doors and elevators can only be activated by additional card readers and fingerprint scanners. All equipment is secured with individually-keyed locks. Some data centres provide security escorts, but this is considered inherently less secure. There is a small risk that security escorts might possibly be distracted or require a bio break, leaving a potential saboteur momentarily unguarded.

In the eventuality of an emergency condition, the data centres are connected to fire, police and other authorities via a remote facility.

This comprehensive approach allows intelligent Internet Data Centers to deliver the ultimate in security-using a range of physical and electronic means -- biometric scanners, 24-hour on-site security guards, bullet-resistant glass, advanced network intrusion detection, and multi-layer firewalls.

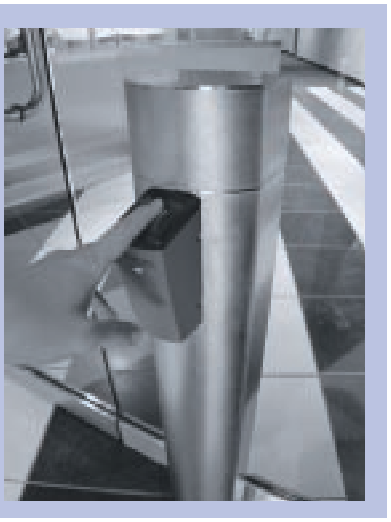

Figure 2 (right): Biometric verification methods ensure that the person is actually who he/she claims to be. Shown here is a fingerprint scanner at a TELUS intelligent Internet Data Centre.

4.4 Network Security

The cornerstone of network security is the firewall, a device that inspects network traffic. Firewalls filter traffic based on packet origination and destination addresses, port numbers and other parameters so that unauthorized traffic can be blocked from entering the data centre's network.

One common network architectural practise is to connect co-location and managed hosting servers on separate subnets. A second layer of firewalling is used to isolate the managed hosting servers from the co-located servers. Sometimes a different vendor of firewall is used for this second layer, so that hackers with great expertise who are familiar with any potential security issues with one firewall product would be stumped by the second product.

Individual data centre customers can also implement a third layer of protection by implementing their own security policies in their own dedicated firewall. Installing, configuring and operating firewalls is one of the most popular professional services offered by data centres.

The next vital IT security measure is an intrusion detection system, or IDS. An IDS monitors the traffic on a network segment to detect potential attack streams.

Many hacker attacks exploit vulnerabilities in communication protocols to deliberately create errors, or other conditions under which mischief can be done. The difference between a firewall and an intrusion detection system is that an IDS maintains protocol state and contextual information in real-time. This allows malicious traffic, which has been passed through the firewall on the basis of using authorized IP addresses and port numbers, to be detected. The IDS can then issue network management alerts, terminate user sessions, or even reconfigure firewalls to block traffic.

In the case of an Internet data centre operated by a carrier, these IT security systems can be integrated or operated independently to offer very high levels of security. For example, most IP backbone operators also have firewall and IDS systems in place.

4.5 Redundancy

Very high availability is almost always achieved through the use of redundancy. In theory, Internet data centres should have no single point of failure. This implies that all systems must have a hot standby available -- even when systems are taken out of service for maintenance.

At minimum, multiple fibre trunks are required, ideally from different service providers and entering the IDC at different physical locations.

The long-term availability of sufficient electricity is a significant consideration. Just consider the recent history of brownouts and blackouts in California, the largest geographical data centre market.

Internet data centres typically use from 85 to 100 watts per square foot to power equipment and meet their air conditioning requirements. This is as much as twenty times the average commercial space usage. When possible, an IDC should have availability to redundant power grids. Uninterruptible power supplies and backup diesel generators are also commonly used to guarantee a continuous supply of electricity.

Given the high power densities of IDCs, a cooling system failure would cause a rapid temperature rise. Although an IDC melt-down has not been publicly reported, it is a real possibility and thus redundant cooling systems are essential. But don't necessarily expect to see two enormous air conditioners in a data centre. This redundancy is sometimes achieved through deliberate over-sizing of a cooling plant composed of modular air conditioners, and the use of redundant cool air distribution facilities, rather than the twinning of a massive single air conditioner.

Figure 3: A single point of ingress and egress not only controls entry authorization, but prevents unauthorized equipment removal and impedes a physical attack.

One concept borrowed directly from corporate data processing centres is the use of raised floors. Although raised floors can simplify cabling, they also provide an efficient air duct for cooling. Some IDC have raised floors that are over two feet high for this reason.

The ultimate redundancy is a second data centre. Some IDC operators are plan to build networks interconnecting their data centres, so that they can remotely back up and restore each other.

TELUS data centres solve this problem in a tidy fashion by using Cisco Systems, Inc.'s global load balancing technology to geographically distribute server loads among data centres. For example, an enterprise could have its web operations hosted simultaneously in both Calgary and Toronto. The load balancing technology senses the user's location and balances the load on a geographical basis for best performance and response time. If the unthinkable were ever to occur and one data centre experienced an outage, the load balancer would automatically shift traffic to the other data centre.

4.6 Management

The availability of information technology skills is another issue driving demand for IDCs. Data centres are typically operated by third parties, an arrangement which offers several benefits.

First, outsourcing web operations to a third party may be a strategic objective, especially when not within the scope of a firm's core business or competencies.

Second, a third party at arm's length can be held accountable for delivering on service level agreements. This might be difficult when using an in-house facility.

And third, a neutral third party can aggregate demand from customers who may not normally wish to interact. A larger scale of operations could benefit them by driving down costs for all involved.

Specialized technical and management expertise is an important component of the data centre value proposition.

5.0 Conclusion

Internet data centres offer a high-availability environment so that mission-critical Internet operations can be offered with minimal risk of interruption.

The high costs of an IDC are usually unaffordable by a single enterprise. But, when shared by many, economies of scale transform a multimillion dollar capital investment into a much lower operating expense. This cost-sharing permits IDC customers to take advantage of risk-reduction technology that would not otherwise be within their reach, and they ultimately obtain much higher reliability at much lower cost.

6.0 Acknowledgements

The contributions of Craig Richardson, Director, Hosting Solutions at TELUS Business Solutions, Calgary in the preparation of this publication are gratefully acknowledged.

About the author

Gregory Graham holds a Bachelor of Electrical Engineering from Concordia University in Montreal and an MBA from the University of Alberta in Edmonton.

He has almost 20 years of experience in the telecommunications and computer industries. His background includes increasingly-responsible roles in marketing and strategic planning at Hewlett-Packard, TELUS Advanced Communications, and Nortel Networks. He was a member of the management team that grew TELUS PLAnet into western Canada's largest Internet service provider, and introduced North America's first residential ADSL service. He also created Canada's first wholesale VoIP call termination service offered by a telco. He can be reached at greg_graham@compuserve.com.