Computer Networks 53 (2009) 2939–2965

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Data center evolution
A tutorial on state of the art, issues, and challenges

Krishna Kant

Intel Corporation, Hillsboro, Oregon, USA

ARTICLE INFO

Keywords:
Data center
Virtualization
InfiniBand
Ethernet
Solid state storage
Power management

ABSTRACT

Data centers form a key part of the infrastructure upon which a variety of information technology services are built. As data centers continue to grow in size and complexity, it is desirable to understand aspects of their design that are worthy of carrying forward, as well as existing or upcoming shortcomings and challenges that would have to be addressed. We envision the data center evolving from owned physical entities to potentially outsourced, virtualized and geographically distributed infrastructures that still attempt to provide the same level of control and isolation that owned infrastructures do. We define a layered model for such data centers and provide a detailed treatment of state of the art and emerging challenges in storage, networking, management and power/thermal aspects.

1. Introduction

Data centers form the backbone of a wide variety of services offered via the Internet including Web-hosting, e-commerce, social networking, and a variety of more general services such as software as a service (SAAS), platform as a service (PAAS), and grid/cloud computing. Some examples of these generic service platforms are Microsoft's Azure platform, Google App engine, Amazon's EC2 platform and Sun's Grid Engine. Virtualization is the key to providing many of these services and is being increasingly used within data centers to achieve better server utilization and more flexible resource allocation. However, virtualization also makes many aspects of data center management more challenging.

As the complexity, variety, and penetration of such services grows, data centers will continue to grow and proliferate. Several forces are shaping the data center landscape and we expect future data centers to be lot more than simply bigger versions of those existing today. These emerging trends – more fully discussed in Section 3 – are expected to turn data centers into distributed, virtualized, multi-layered infrastructures that pose a variety of difficult challenges.

In this paper, we provide a tutorial coverage of a variety of emerging issues in designing and managing large virtualized data centers. In particular, we consider a layered model of virtualized data centers and discuss storage, networking, management, and power/thermal issues for such a model. Because of the vastness of the space, we shall avoid detailed treatment of certain well researched issues. In particular, we do not delve into the intricacies of virtualization techniques, virtual machine migration and scheduling in virtualized environments.

The organization of the paper is as follows. Section 2 discusses the organization of a data center and points out several challenging areas in data center management. Section 3 discusses emerging trends in data centers and new issues posed by them. Subsequent sections then discuss specific issues in detail including storage, networking, management and power/thermal issues. Finally, Section 8 summarizes the discussion.

E-mail address: krishna.kant@intel.com

1389-1286/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.comnet.2009.10.004

2. Data center organization and issues

2.1. Rack-level physical organization

A data center is generally organized in rows of "racks" where each rack contains modular assets such as servers, switches, storage "bricks", or specialized appliances as shown in Fig. 1. A standard rack is 78 in. high, 23–25 in. wide and 26–30 in. deep. Typically, each rack takes a number of modular "rack mount" assets inserted horizontally into the racks. The asset thickness is measured using an unit called "U", which is 45 mm (or approximately 1.8 in.). An overwhelming majority of servers are single or dual socket processors and can fit the 1U size, but larger ones (e.g., 4-socket multiprocessors) may require 2U or larger sizes. A standard rack can take a total of 42 1U assets when completely filled. The sophistication of the rack itself may vary greatly – in the simplest case, it is nothing more than a metal enclosure. Additional features may include rack power distribution, built-in KVM (keyboard–video–mouse) switch, rack-level air or liquid cooling, and perhaps even a rack-level management unit.

For greater compactness and functionality, servers can be housed in a self-contained chassis which itself slides into the rack. With 13 in. high chassis, six chassis can fit into a single rack. A chassis comes complete with its own power supply, fans, backplane interconnect, and management infrastructure. The chassis provides standard size slots where one could insert modular assets (usually known as *blades*). A single chassis can hold up to 16 1U servers, thereby providing a theoretical rack capacity of 96 modular assets.

The substantial increase in server density achievable by using the blade form factor results in corresponding increase in per-rack power consumption which, in turn, can seriously tax the power delivery infrastructure. In particular, many older data centers are designed with about 7 KW per-rack power rating, whereas racks loaded with blade servers could approach 21 KW. There is a similar issue with respect to thermal density – the cooling infrastructure may be unable to handle the offered thermal load. The net result is that it may be impossible to load the racks to their capacity. For some applications, a fully loaded rack may not offer the required peak network or storage bandwidth (BW) either, thereby requiring careful management of resources to stay within the BW limits.

Fig. 1. Physical organization of a data center.

2.2. Storage and networking infrastructure

Storage in data centers may be provided in multiple ways. Often the high performance storage is housed in special "storage towers" that allow transparent remote access to the storage irrespective of the number and types of physical storage devices used. Storage may also be provided in smaller "storage bricks" located in rack or chassis slots or directly integrated with the servers. In all cases, an efficient network access to the storage is crucial.

A data center typically requires four types of network accesses, and could potentially use four different types of physical networks. The client–server network provides external access into the data center, and necessarily uses a commodity technology such as the wired Ethernet or wireless LAN. Server-to-server network provides high-speed communication between servers and may use Ethernet, InfiniBand (IBA) or other technologies. The storage access has traditionally been provided by Fiber Channel but could also use Ethernet or InfiniBand. Finally, the network used for management is also typically Ethernet but may either use separate cabling or exist as a "sideband" on the mainstream network.

Both mainstream and storage networks typically follow identical configuration. For blade servers mounted on a chassis, the chassis provides a switch through which all the servers in the chassis connect to outside servers. The switches are duplexed for reliability and may be arranged for load sharing when both switches are working. In order to keep the network manageable, the overall topology is basically a tree with full connectivity at the root level. For example, each chassis level (or level 1) switch has an *uplink* leading to the level 2 switch, so that communication between two servers in different chassis must go through at least three switches. Depending on the size of the data center, the multiple level 2 switches may be either connected into a full mesh, or go through one or more level 3 switches. The biggest issue with such a structure is potential bandwidth inadequacy at higher levels. Generally, uplinks are designed for a specific *oversubscription ratio* since providing a full bisection bandwidth is usually not feasible. For example, 20 servers, each with a 1 GB/s Ethernet may share a single 10 GB/s Ethernet uplink for a oversubscription ratio of 2.0. This may be troublesome if the workload mapping is such that there is substantial non-local communication. Since storage is traditionally provided in a separate storage tower, all storage traffic usually crosses the chassis uplink on the storage network. As data centers grow in size, a more scalable network architecture becomes necessary.

2.3. Management infrastructure

Each server usually carries a management controller called the BMC (baseboard management controller). The management network terminates at the BMC of each server. When the management network is implemented as a "sideband" network, no additional switches are required for it; otherwise, a management switch is required in each chassis/rack to support external communication. The basic functions of the BMC include monitoring of various hardware sensors, managing various hardware and software alerts, booting up and shutting down the server, maintaining configuration data of various devices and drivers, and providing remote management capabilities. Each chassis or rack may itself sport its own higher level management controller which communicates with the lower level controller.

Configuration management is a rather generic term and can refer to management of parameter settings of a variety of objects that are of interest in effectively utilizing the computer system infrastructure from individual devices up to complex services running on large networked clusters. Some of this management clearly belongs to the baseboard management controller (BMC) or corresponding higher level management chain. This is often known as *out-of-band* (OOB) management since it is done without involvement of main CPU or the OS. Other activities may be more appropriate for *in-band* management and may be done by the main CPU in hardware, in OS, or in the middleware. The higher level management may run on separate systems that have both in-band and OOB interfaces. On a server, the most critical OOB functions belong to the preboot phase and in monitoring of server health while the OS is running. On other assets such as switches, routers, and storage bricks the management is necessarily OOB.

2.4. Electrical and cooling infrastructure

Even medium-sized data centers can sport peak power consumption of several megawatts or more. For such power loads, it becomes necessary to supply power using high voltage lines (e.g., 33 KV, 3 phase) and step it down on premises to the 280–480 V (3 phase) range for routing through the uninterrupted power supply (UPS). The UPS unit needs to convert AC to DC to charge its batteries and then convert DC to AC on the output end. Since the UPS unit sits directly in the power path, it can continue to supply output power uninterrupted in case of input power loss. The output of UPS (usually 240/120 V, single phase) is routed to the power distribution unit (PDU) which, in turn, supplies power to individual rack-mounted servers or blade chassis. Next the power is stepped down, converted from AC to DC, and partially regulated in order to yield the typical ± 12 and ± 5 V outputs with the desired current ratings (20–100 A). These voltages are delivered to the motherboard where the voltage regulators (VRs) must convert them to as many voltage rails as the server design demands. For example, in an IBM blade server, the supported voltage rails include 5–6 V (3.3 V down to 1.1 V), in addition to the 12 V and 5 V rails.

Each one of these power conversion/distribution stages results in power loss, with some stages showing efficiencies in 85–95% range or worse. It is thus not surprising that the cumulative power efficiency by the time we get down to voltage rails on the motherboard is only 50% or less (excluding cooling, lighting, and other auxiliary power uses). Thus there is a significant scope for gaining power efficiencies by a better design of power distribution and conversion infrastructure.

The cooling infrastructure in a data center can be quite elaborate and expensive involving building level air-conditioning units requiring large chiller plants, fans and air recirculation systems. Evolving cooling technologies tend to emphasize more localized cooling or try to simplify cooling infrastructure. The server racks are generally placed on a raised plenum and arranged in alternately back-facing and front-facing aisles as shown in Fig. 2. Cold air is forced up in the front facing aisles and the server or chassis fans draw the cold air through the server to the back. The hot air on the back then rises and is directed (sometimes by using some deflectors) towards the chiller plant for cooling and recirculation. This basic setup is not expensive but can also create hot spots either due to uneven cooling or the mixing of hot and cold air.

2.5. Major data center issues

Data center applications increasingly involve access to massive data sets, real-time data mining, and streaming media delivery that place heavy demands on the storage

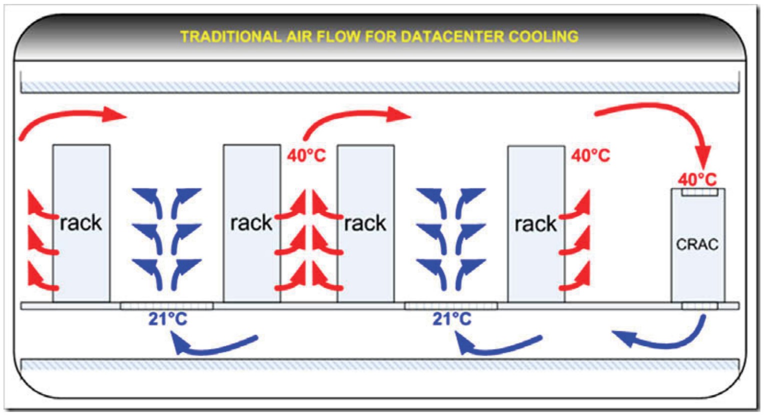

Fig. 2. Cooling in a data center.

infrastructure. Efficient access to large amounts of storage necessitates not only high performance file systems but also high performance storage technologies such as solid-state storage (SSD) media. These issues are discussed in Section 5. Streaming large amounts of data (from disks or SSDs) also requires high-speed, low-latency networks. In clustered applications, the inter-process communication (IPC) often involves rather small messages but with very low-latency requirements. These applications may also use remote main memories as "network caches" of data and thus tax the networking capabilities. It is much cheaper to carry all types of data – client–server, IPC, storage and perhaps management – on the same physical fabric such as Ethernet. However, doing so requires sophisticated QoS capabilities that are not necessarily available in existing protocols. These aspects are discussed in Section 4.

Configuration management is a vital component for the smooth operation of data centers but has not received much attention in literature. Configuration management is required at multiple levels, ranging from servers to server enclosures to the entire data center. Virtualized environments introduce issues of configuration management at a logical – rather than physical – level as well. As the complexity of servers, operating environments, and applications increases, effective real-time management of large heterogeneous data centers becomes quite complex. These challenges and some approaches are discussed in Section 6.

The increasing size of data centers not only results in high utility costs [1] but also leads to significant challenges in power and thermal management [82]. It is estimated that the total data center energy consumption as a percentage of total US energy consumption doubled between 2000 and 2007 and is set to double yet again by 2012. The high utility costs and environmental impact of such an increase are reasons enough to address power consumption. Additionally, high power consumption also results in unsustainable current, power, and thermal densities, and inefficient usage of data center space. Dealing with power/thermal issues effectively requires power, cooling and thermal control techniques at multiple levels (e.g., device, system, enclosure, etc.) and across multiple domains (e.g., hardware, OS and systems management). In many cases, power/thermal management impacts performance and thus requires a combined treatment of power and performance. These issues are discussed in Section 7.

As data centers increase in size and criticality, they become increasingly attractive targets of attack since an isolated vulnerability can be exploited to impact a large number of customers and/or large amounts of sensitive data [14]. Thus a fundamental security challenge for data centers is to find workable mechanisms that can reduce this growth of vulnerability with size. Basically, the security must be implemented so that no single compromise can provide access to a large number of machines or large amount of data. Another important issue is that in a virtualized outsourced environment, it is no longer possible to speak of "inside" and "outside" of data center – the intruders could well be those sharing the same physical infrastructure for their business purposes. Finally, the basic virtualization techniques themselves enhance vulnerabilities since the flexibility provided by virtualization can be easily exploited for disruption and denial of service. For example, any vulnerability in mapping VM level attributes to the physical system can be exploited to sabotage the entire system. Due to limited space, we do not, however, delve into security issues in this paper.

3. Future directions in data center evolution

Traditional data centers have evolved as large computational facilities solely owned and operated by a single entity – commercial or otherwise. However, the forces in play are resulting in data centers moving towards much more complex ownership scenarios. For example, just as virtualization allows consolidation and cost savings within a data center, virtualization across data centers could allow a much higher level of aggregation. This notion leads to the possibility of "out-sourced" data centers that allows an organization to run a large data center without having to own the physical infrastructure. Cloud computing, in fact, provides exactly such a capability except that in cloud computing the resources are generally obtained dynamically for short periods and underlying management of these resources is entirely hidden from the user. Subscribers of virtual data centers would typically want longer-term arrangements and much more control over the infrastructure given to them. There is a move afoot to provide *Enterprise Cloud* facilities whose goals are similar to those discussed here [2]. The distributed virtualized data center model discussed here is similar to the one introduced in [78].

In the following we present a 4-layer conceptual model of future data centers shown in Fig. 3 that subsumes a wide range of emergent data center implementations. In this depiction, rectangles refer to software layers and ellipses refer to the resulting abstractions.

The bottom layer in this conceptual model is the *Physical Infrastructure Layer* (PIL) that manages the physical infrastructure (often known as "server farm") installed in a given location. Because of the increasing cost of the power consumed, space occupied, and management personnel required, server farms are already being located closer to sources of cheap electricity, water, land, and manpower. These locations are by their nature geographically removed from areas of heavy service demand, and thus the developments in ultra high-speed networking over long distances are essential enablers of such remotely located server farms. In addition to the management of physical computing hardware, the PIL can allow for larger-scale consolidation by providing capabilities to carve out well-isolated sections of the server farm (or "server patches") and assign them to different "customers." In this case, the PIL will be responsible for management of boundaries around the server patch in terms of security, traffic firewalling, and reserving access bandwidth. For example, set up and management of virtual LANs will be done by PIL.

The next layer is the *Virtual Infrastructure Layer* (VIL) which exploits the virtualization capabilities available in individual servers, network and storage elements to support the notion of a *virtual cluster*, i.e., a set of virtual or real nodes along with QoS controlled paths to satisfy their

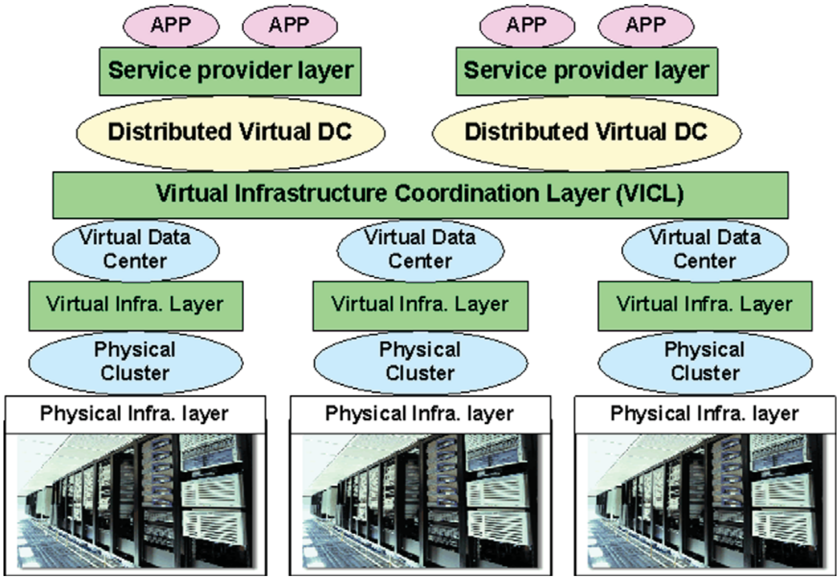

Fig. 3. Logical organization of future data centers.

communication needs. In many cases, the VIL will be internal to an organization who has leased an entire physical server patch to run its business. However, it is also conceivable that VIL services are actually under the control of infrastructure provider that effectively presents a *virtual server patch* abstraction to its customers. This is similar to cloud computing, except that the subscriber to a virtual server patch would expect explicit SLAs in terms of computational, storage and networking infrastructure allocated to it and would need enough visibility to provide its own next level management required for running multiple services or applications.

The third layer in our model is the *Virtual Infrastructure Coordination Layer* (VICL) whose purpose is to tie up virtual server patches across multiple physical server farms in order to create a geographically distributed virtualized data center (DVDC). This layer must define and manage virtual pipes between various virtual data centers. This layer would also be responsible for cross-geographic location application deployment, replication and migration whenever that makes sense. Depending on its capabilities, VICL could be exploited for other purposes as well, such as reducing energy costs by spreading load across time-zones and utility rates, providing disaster or large scale failure tolerance, and even enabling truly large-scale distributed computations.

Finally, the *Service Provider Layer* (SPL) is responsible for managing and running applications on the DVDC constructed by the VICL. The SPL would require substantial visibility into the physical configuration, performance, latency, availability and other aspects of the DVDC so that it can manage the applications effectively. It is expected that SPL will be owned by the customer directly.

The model in Fig. 3 subsumes everything from a non-virtualized, single location data center entirely owned by a single organization all the way up to a geographically distributed, fully virtualized data center where each layer possibly has a separate owner. The latter extreme provides a number of advantages in terms of consolidation, agility, and flexibility, but it also poses a number of difficult challenges in terms of security, SLA definition and enforcement, efficiency and issues of layer separation. For this reason, real data centers are likely to be limited instances of this general model.

In subsequent sections, we shall address the needs of such DVDC's when relevant, although many of the issues apply to traditional data centers as well.

4. Data center networking

4.1. Networking infrastructure in data centers

The increasing complexity and sophistication of data center applications demands new features in the data center network. For clustered applications, servers often need to exchange inter-process communication (IPC) messages for synchronization and data exchange, and such messages may require very low-latency in order to reduce process stalls. Direct data exchange between servers may also be motivated by low access latency to data residing in the memory of another server as opposed to retrieving it from the local secondary storage [18]. Furthermore, mixing of different types of data on the same networking fabric may necessitate QoS mechanisms for performance isolation. These requirements have led to considerable activity in the design and use of low-latency specialized data center fabrics such as PCI-Express based backplane interconnects, InfiniBand (IBA) [37], data center Ethernet [40,7], and lightweight transport protocols implemented directly over the Ethernet layer [5]. We shall survey some of these developments in subsequent sections before examining networking challenges in data centers.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase® Smarter legal research.