WHITE PAPER

Introduction to Cloud Design

Four Design Principals For IaaS

What is a Cloud

Cloud computing is a collection of technologies and practices used to abstract the provisioning and management of computer hardware. The goal is to simplify the users experience so they can get the benefit of compute resources on demand; or in the language of cloud computing "as a service".

The resources that comprise an individual cloud are generally made available to end users using an interface that conforms to one of three levels of abstractions. From most granular to most abstract these are Infrastructure-as-a-Service (Iaas), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS).

Applications, offered to end users through a web browser or thin client, who are almost entirely stored managed and updated in the cloud, are termed Software-as-a-Service (SaaS). Deploying an application using a SaaS model is typical for social networking, collaboration, media and content management types of applications. However we are seeing more traditional desktop applications moving to this model. Examples are far ranging and include CRM's like salesforce.com, email like gmail and web based games like FarmVille.

Platform-as-a-Service (PaaS) is a level of abstraction below SaaS. The Platform is a tailored environment meant to expedite the development of applications. The platform contains a specification of compute power for hosting some aspects of the application as well as a set of software like web servers, databases, load balancers etc. These pre-build components of the platform make the deployment and maintenance of these resources simple by moving the responsibility off the application developer and to the cloud maintainer. PaaS provides turn-key deployment of common platform elements like apache web-servers, SQL databases, load-balancers etc.

Infrastructure-as-a-service (IaaS) is the creation of virtual hardware resources including virtual machines, virtual networks and virtualized storage. IaaS is very tightly coupled with virtualization concepts and typically forms the basis for the higher level abstractions of PaaS and Saas.

Although storage is logically a part of IaaS, it is often considered separately. This is particularly the case where cloud storage is becoming more commonly used without other parts of the IaaS. Examples include Amazon S3 and Rackspace Cloud Files as well as SaaS applications like dropbox and box.com.

This paper outlines four principal design considerations for IaaS cloud deployments and examines the benefits of Mellanox's Ethernet and InfiniBand interconnects for IaaS.

Why Mellanox for the Cloud

Mellanox Technologies is the world's leader in RDMA technology. RDMA (Remote Dynamic Memory Access) is a network adapter capability allowing Memory to Memory transfers between networks connected commodity x86 systems. Mellanox RDMA allows the movement of network and storage data between virtual machines, with the highest bandwidth, lowest latency and least CPU cycles of any other interconnect technology on the market. Mellanox's interconnect technologies primarily benefit cloud providers who require to build scalable and low-cost IaaS clouds. This paper focuses only on benefits to IaaS, but since SaaS and PaaS are layers on top of an IaaS foundation, benefits are inherited.

Design Considerations in Building an IaaS Cloud

Through significant customer engagements, building data centers and working closely with IaaS architects and administrations, Mellanox observed a few facts that all of the best cloud data centers in the world share. These similarities are segmented into the four design principles outlined below:

Principle #1 – Scalable Physical Design

The best IaaS clouds have quick and easy ways to deploy physical machines and to begin using them as part of the virtual infrastructure. Most public and private clouds will start small and need to add additional capacity as more users start to bring their applications to the cloud.

The primary goal is to be able to use a highly dense physical form factor. A second goal is to have a highly modular design. These are interrelated goals because they allow new resources to be added to the cloud in a granular fashion keeping provisioning simple. Dense form factors are generally more efficient in terms of heat and power yielding better TCO. Generally the purchase price for high density nodes is on par or cheaper than compared to multi-rack unit equivalents. The two constraints that prevent IaaS architects from deploying high density solutions are typically storage capacity of these models and IO expandability.

During a new deployment of physical systems to the existing cloud the goal is to prevent is introducing down-time in the existing system. This happens because complexity in physical installation of the new resources. When designing a cloud cluster one important consideration is to avoid performance reduction due to rebalancing once the physical resources are installed.

From the networking perspective it is preferred to choose a technology and a topology that allows use of dense form factors and minimizes difficulty of physical installation. Using a high bandwidth interconnect solution enable a design that maximize the number of virtual infrastructure devices over the fewest physical connections. Further it provides a converge different traffic types (i.e. Management, Storage, Network) over a single wire, simplifying the physical installation.

Assume that a given virtual infrastructure device needs approximately 1Gb/s of connectivity including network, storage and management traffic which is probably conservative. This number is particularly conservative if the IaaS is designed to decouple storage resources (principal 3) from the virtual machine hosts. Based on these assumptions approximately n gbps of bandwidth is required where n is the number of virtual machines running on any given VM host. Unless the IaaS is designed to run scientific or rendering types of compute jobs with Nehalem or later components it is expected a typical cloud deployments will run anywhere from 15-30 VMs per physical infrastructure server. The only technology's that allow this measure of bandwidth over a single port are 40Gb/s Ethernet or InfiniBand at 40-56Gb/s.

Having a solution that provides the needed capacity in a single port is important because it allows for a dense design. As previously discussed this benefits both capital and operation expenses. Additionally it simplifies configuration by eliminating the need to bond links both for aggregation and redundancy.

An ideally form factor for this type of dense networking puts the NIC directly on the motherboard (LOM) or in a specialized add on card (Mezz or ALOM). Examples include the Hewlett-Packard SL 390 which has on motherboard 40 Gigabit InfiniBand or the Dell C6100 which has a Mezzanine add on option for Mellanox 40Gb/s Ethernet or InfiniBand. In both examples the PCIe slot in the system remains open allowing for expansion, despite the very dense form factor.

Principle #2 – Simple Provisioning Rules

The best IaaS clouds have quick and easy ways to deploy new virtual machines and to connect them to the rest of the virtual infrastructure. This task usually consists of identifying the right hypervisor to create the virtual machine, provisioning the storage volume required for the virtual machine and then provisioning the network connectivity required by the VM. Scheduling is the act of mapping virtual infrastructure to physical resources. It is also the most challenging aspects of cloud design. Having a simple method for provisioning means making the scheduling choices limited, inconsequential or automated.

The decision to automate or manually carry out provisioning in a cloud largely depends on the size of the deployment, the customer base and the degree to which allocations are dynamic or static. The best solutions have the ability to work in either fashion using a well-integrated central management system.

Mellanox's Unified Fabric Management (UFM) software provides this type of solution. UFM gives the administrators control over all aspects of the nodes including provisioning, QoS and vNIC management, as well as ability to remotely manage firmware and software revisions for the network. UFM also provides simplified automation for common tasks. When changes occur in the network such as VM migrations UFM will automatically provision the physical and virtual networking elements to keep connectivity and policy intact. UFM takes care of the network scheduling and integrates with a range of industry standard automated schedulers from MOAB to OpenStack to handle the non-network scheduling.

Most IaaS solutions today have the ability to enforce SLAs for components like CPU and Memory. However only the best IaaS deployments consider and plan for providing SLAs that cover networking in the cloud. UFM has a unique performance monitoring engine, that can provide near real-time information on performance volumes and potential issues such as traffic bottlenecks. The information is automatically correlated to UFM's Logical Model that represents the cloud applications and services. This enables the user to finally see have one place the ability to see what is the service level the cloud services get on the network, their bandwidth consumption and any traffic issues they have, together with the tweaking and provisioning capabilities – so, a full operational cycle is completed.

The InfiniBand technology also offers some key advantages that cloud vendors are taking advantage of. InfiniBand is a self healing fabric, meaning that central management is constantly and automatically monitoring for link failures and congested links in the network. The technology has standard based methods to dynamically balance or correct routing in the network based on real-time information. Secondly InfiniBand has an advantage of scale, where single subnets typically grow to include 1000's of nodes, in comparison to Ethernets hundreds. This simplifies the network and allows for reduced complexity and errors associated with router configuration. Finally InfiniBand offers lossless communication which is critical to support storage over network protocols and to avoid network congestion.

Principle #3 – An Elastic Design

An IaaS system should have the ability to adapt to changing load requirements. To the IaaS should have the ability to scale-up when load increases and to consolidate when load decreases. Work-loads may also change dynamically and being able to rebalance is important. From the IaaS perspective this means being able to readjust the physical location of virtual machines or storage volumes to balance load on storage subsystems.

The single most important design decision a cloud infrastructure manager can make is to decouple the storage nodes from the nodes running the VMs. This creates three big advantages. Most importantly it allows efficient allocation of the storage. When storage is clustered it will not become fragmented. Secondly it means that the scheduling for running VM images is never limited to the local storage subsystem of a node. This makes allocation simple. Finally it allows greater elasticity because live migration is only of a running image and never includes the storage volume.

The challenge of providing decoupled storage is network performance. Running remote file system or block based protocols like NFS or iSCSI will require additional bandwidth per VM. However bandwidth isn't enough, because the storage is remote access times will suffer unless a low-latency network is used. Distributed file systems are also popular for this task as they spread the bandwidth requirement between many nodes. However this additional communication will create additional load across the network. To mitigate this effect without creating an entirely separate storage network you need to design with a network that has strong mechanisms for traffic isolation and QoS rules.

The final risk when decoupling the storage from the VM infrastructure node is increased CPU utilization for processing storage traffic. Using a network adapter that provides offload or better yet RDMA for moving storage traffic solves this issue.

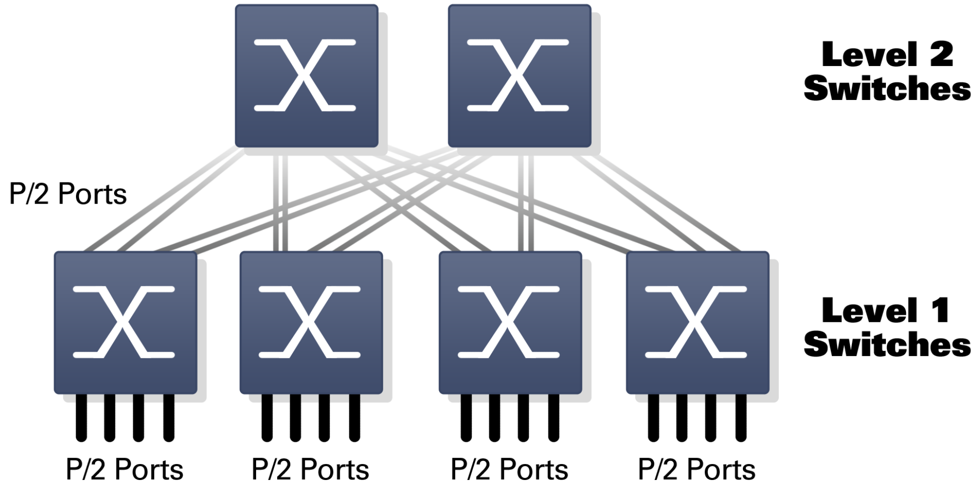
Once storage has been decoupled from the live migration process, the cloud has the ability to be far more agile in redistributing loads between infrastructure nodes. The final remaining challenge is to minimize migration time which is a high CPU, high network bandwidth operation. To do this having a low latency, high bandwidth interconnect such as 40Gb/s Ethernet or 56Gb/s InfiniBand makes VM live migration up to 4 times faster than on traditional 10GbE networks. (Beck, 2011)

Principle #4 – Capacity to Support East-West Traffic Flows

North-south traffic is defined as the flow from the hypervisor through a top of rack switch and then an aggregation switch out to the internet. This is typically traffic generated from the end user or the WAN. However studies have shown that the majority of traffic in the cloud is east west traffic, meaning between virtual servers within the cloud. This is mostly due to the standard composition of a SaaS application which will employ many virtual infrastructure and platform entities to build an application capable of providing dynamic content within an expected response time. Platform elements like memcached servers, load balancers and databases must retrieve data from infrastructure entities like storage and cloud files, process and aggregate this information. (Morgan, 2011) All this processing happens internally to the cloud generating very large east-west traffic volumes relative to the final North bound responds to the WAN.

"Wrong" scheduling decisions with regard to east-west traffic are generally made when the cloud infrastructure is unbalanced and the resources are fragmented across the physical infrastructure. Without knowing ahead of time which virtual infrastructure devices will need to interact the only way to guarantee the east-west traffic flows will function correctly is to use a network topology that guarantees constant cross sectional bandwidth and latency such as a "Fat-Tree" architecture.

Fat-Tree architectures differ from a traditional 3-tier data center architecture in that the Top of Rack (TOR) and the aggregation layers are designed in a fully interconnect manner. The tree will either be non-blocking or have a low blocking ratio. For hyperscale deployments an aggregated FAT-Trees can be bound within a POD and aggregated to a third tier. However in this configuration applications of clients should be bound within a single POD to make sure they can take advantage of the fat-tree.



Summary

By applying the four principals recommended above in an IaaS design, a highly optimized and efficient cloud deployment can be achieved. This methodology enables the best hardware usage, by allowing, a highly dynamic IaaS environment. Mellanox's hardware and software products inherently supports these principals offering the best ROI and TCO:

Principle #1 – Scalable Physical Design

- Choose a Dense and Modular Form Factor with Networking on board (or Mezz)
- Converge Network, Storage and Management IO over a Single Wire

Principle #2 – Simple Provisioning Rules

- Use Tools that allow both automation and manual provision
- Consider InfiniBand provides dynamic routing and is self-healing

Principle #3 – An Elastic Design

- Decouple your storage nodes from your infrastructure nodes
- Use a high-bandwidth interconnect to reduce live migration times by 4x

Principle #4 – Plan Capacity to Support East-West Traffic Flows

- Use Fat-Tree topology to avoid east-west traffic congestion
- Consider InfiniBand which scales to thousands of nodes in a single L2

Feature	Mellanox FDR (56Gb/s) InfiniBand	Mellanox 10/40Gb/s Ethernet	Traditional 1/10Gb/s Ethernet
Low Latency	Lowest	Low – VMA or RoCE	High
Lossless Mode of Operation	Native	Yes, with DCB	Yes, with DCB
Centralized Management	Yes	Yes	No
Support for Automated Schedulers	Yes	Yes	Maybe
Support for RDMA Storage	Yes	Yes	Maybe
Hardware Enforced Segment Isolation	Yes	Yes	No
Self-Healing	Yes	No	No
Dynamic Load Balance	Yes	No	No

Works Cited

Beck, M. (2011). VM Migration Acceleration over 40GbE.

Retrieved from http://www.mellanox.com/pdf/PPT/VM%20Migration%20over%2040GigE.pdf

Morgan, T. P. (2011, 11 29). The Register. Retrieved from The Register:

w://www.theregister.co.uk/2011/11/29/cisco_cloud_data_center_traffic_index/

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com