ascent
Thought leadership from Atos

white paper

Fabric-based Computing

Your business technologists. Powering progress

Atos

At the heart of today's datacenters are cables – miles and miles of cables – which connect islands of computer power to islands of storage capacity. These cables impose bandwidth, flexibility and management constraints that inhibit truly scalable datacenter resources to meet the new challenges presented by Big Data, Pattern-Based Strategy, Social Collaboration and other IT challenges.

Efficient resource utilization and automated provisioning with multi-tenancy has been promised by virtualization and cloud technologies for years now, but today's virtualization platforms remain islands of resource pools.

Fabric-based Computing is a new paradigm, meshing interconnections of IT Datacenter resources via high bandwidth links to become a computing "fabric".

What for?

- Service-specific server architectures are no longer hard-wired into the physical infrastructure. Instead, the server landscape becomes fully software defined.
- Re-purposing of resources can be done at the click of a mouse or even automatically according to demand, allocating and reallocating resources across workloads, unencumbered by the physical infrastructure.
- Changes are executed in "near time", allowing Datacenter Infrastructure Management (DCIM) tools to dynamically adjust allocation of resources (CPU, memory, I/O, storage, cooling, etc.) to optimize e.g. power consumption.
- Dynamic resource allocation for services can be automated based on agreed SLAs, user experience or job execution time.

Published in August 2013

Fabric-based Computing

Contents

About the Authors
This white paper was edited by Nicolas Roux, Purchasing Director and member of the Atos Scientific Community (nicolas.roux@atos.net), based on contributions from Antoine Fressancourt, Thomas Hoberg, Martin Pfeil, Guy Lidbetter, Adam Dolman, Nicolas Roux, Boris Scharinger, Peter Joris, Tom Coenraad, Alexandro Caballero, Jean-Francois James. Thanks to Anna Maria Juan Ferrer and Paul Albada Jelgersma for their kind inputs, contributions and detailed review.

About the Atos Scientific Community
The Atos Scientific Community is a network of some 100 top scientists, representing a mix of all skills and backgrounds, and coming from all geographies where Atos operates. Publicly launched by Thierry Breton, Chairman and CEO of Atos, the establishment of this community highlights the importance of innovation in the dynamic IT services market and the need for a proactive approach to identify and anticipate game changing technologies.

General Purpose and Specific Purpose Datacenters

General Purpose Datacenters

Datacenter infrastructure comprises the full stack of power, cooling, network, computing power, storage, and applications necessary to support enterprise business. It is usually consolidated in physical hosting facilities providing optimized space, power supply, cooling and security.

Datacenters operated by a single business with a broad variety of applications or by an IT service provider, serving many customers, are classified as General Purpose. They need to support practically everything, any type, any load and any customer with maximum flexibility and minimal lead-time. That type of flexibility comes at the cost of lower run-time efficiency.

Specific Purpose Datacenters

University and military research, scientific laboratories, financial institutions or search engine providers have specific usage patterns and require very scalable Datacenter infrastructures tailored for a single specific purpose. Grid computing, a federation of clusters combining CPUs and high-speed network interconnects with software into a unified high performance system, is often used in this type of environment: they could scale out to hundreds or thousands of similar types of nodes working together as one integrated system.

The special purpose Datacenter offers a high degree of scale out flexibility, but it is restricted when it comes to functional diversity. In fact the essential functions such as resource or workload management are performed by the integrated nodes and their application layer and not by an independent separate Datacenter management layer.

The Commercial Challenge

Over the last 20 years the Datacenter business has evolved from a Real Estate segment to an added-value business where the most competitive players show their strengths in their ability to provide integral solutions. Although Datacenter management generates limited margins, the creation of a Datacenter infrastructure still requires significant investments. Increasing the average utilization, which translates into scaling effects, is the primary commercial driver. Resource silos as seen in many of today's General Purpose Datacenters must be overcome.

Fabric-based Computing is an approach that aims to take the run-time efficiency of Specific Purpose Datacenters (with their high utilization, scaling effects and a full resource management/optimization pool) to the General Purpose Datacenter without losing their primary benefits; the maximum flexibility and minimal lead-time.

Fabric-based Datacenters carefully exploit the best practices, the compliance standards and anticipate the customer needs, to offer domain-specific but at the same time pre-integrated building blocks, which can be easily repurposed and quickly deployed.

Definition of Fabric-based Computing

Fabrics are used to aggregate components into a system. For example, a Fiber Channel fabric aggregates servers and storage to form pools of resources that enable applications or tasks to share the pooled resources.

Over the last 20 years, the technology used to build networks has not really changed: bridges, switches and routers, usually interconnected to handle "vertical" North-South traffic flowcharts that are typical of the current branch or campus networks. Data flows from one device, then down to a Datacenter or server, and up again to another device. But Datacenter traffic is more horizontal – flowing East-West directly from one 'data processing' entity to another.

Fabric-based Datacenters intend to provide a flat, scalable, multipurpose, any-to-any architecture, where each device has a direct connection to any other.

Gartner introduced "Fabric-Based Infrastructure" as the overarching term for the emerging technology trends affecting the Datacenter architecture. These trends all focus on problems across multiple scales that inhibit full-scale virtualization of the entire Datacenter. Gartner also defines a "Fabric-Based Computer" that virtualizes CPU, memory, IO bus and offload processors (such as graphics processors) into individual resource pools, where any of them can scale independent of the other.

Figure 1: Traditional Datacenter Network Layouts

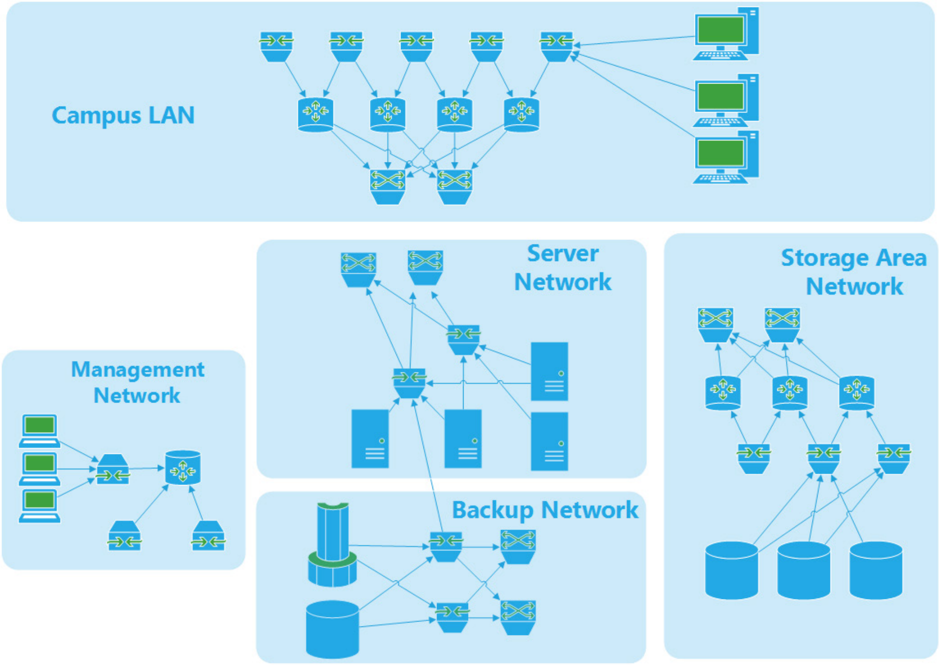
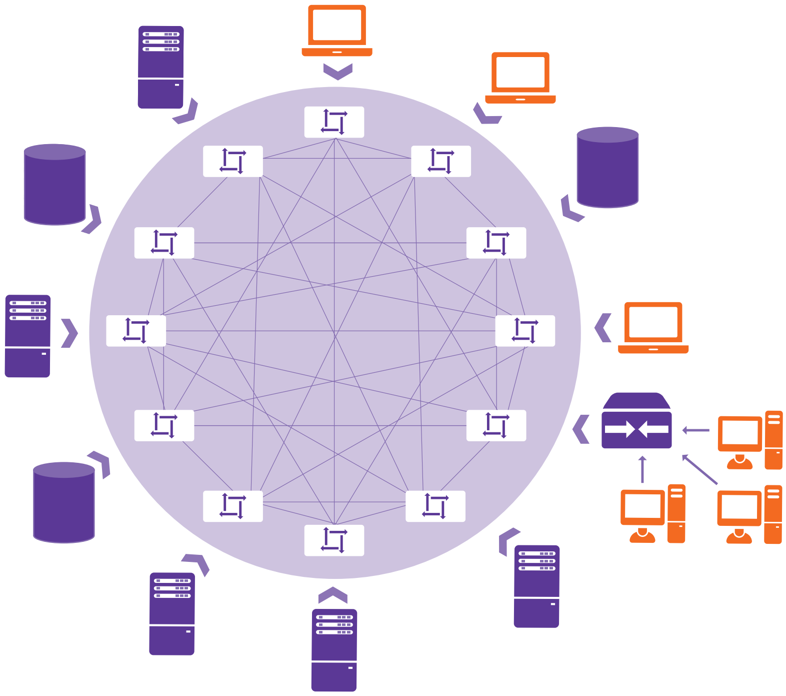


Figure 2: Fabric-based Datacenter Infrastructure



Enablers and Inhibitors

The following section describes a number of technology trends and their impact on Fabric-based Datacenters.

We identify different typologies:

- **Enablers** such as virtualization, component commoditization, Runbook Automation, Convergence and general purpose Vertically Integrated Stacks
- **Supplements** such as Datacenter Infrastructure Management
- **Inhibitors** such as application-specific Vertically Integrated Stacks and In Memory Computing.

The Enabling Role of Virtualization

Generally speaking, to virtualize a set of diverse concrete resources is to access them through a single uniform interface that, from the upper layer perspective, enables them to behave as one unified resource that can be shared with multiple degrees of dynamic behavior. Virtualization brings abstraction and creates a single pool of resources illusion.

Virtualization solutions improve the use of resources by more dynamically distributing workloads across server, storage and network, increasing the utilization rates to 70-80%. However, the benefits of virtualization go far beyond increased utilization. Virtualization allows workload management of physical resources by management of logical software components. As a result, virtualization adds many additional benefits like increased availability, advanced automation and relative independence from the underlying hardware layer.

For the sake of simplicity, we distinguish here the following types of virtualization:

- Bare Metal Virtualization or Hardware Virtualization (aka Hypervisor Type 1). A bare metal environment is a computer system or network in which one or some virtual machine is installed directly on hardware rather than within a host operating system (OS)
- OS Level Virtualization (aka Hypervisor Type 2). Hosted hypervisors that run on top of a conventional operating system environment. With the hypervisor layer as a distinct second software level, guest operating systems run at the third level above the hardware
- Application Virtualization. Software technology that enables and provides isolation and encapsulation at application level.

However, virtualization also adds complexity:

- The physical layer still has to be managed, thus adding constraints like consistency requirements and compatibility matrixes across the different physical components.
- The management of the virtualization stack itself comes in addition to the management of the rest of the infrastructure – and it requires specific skills and procedures.
- Virtualization often implies strong interactions between the IT resources and therefore requires finely tuned and optimized tuning of these resources.

The more mainstream virtualization solutions are becoming, the more vendors are tightly integrating them into their offers. Beyond integration, the ability to provide an abstraction layer that provides an entry point for automation software is key. Otherwise, standardization would be sufficient to avoid the need for an abstraction layer.

Virtualization solutions and their inherent capacity to manage and balance workloads across a large number of nodes will be very important to handle an over increasing number of components composing of an instance – thousands of cores, ports, spindles. This means both more flexibility will large pools of resources and more segmentation with the ability to allocate precise workloads.

Virtualization is a key enabler of software-managed environments, which inherently provides highly automated, orchestrated and flexible infrastructure services.

Commoditization of Components

Commoditization is a major trend in the IT industry strongly driven by cost optimization, but it also faces resistance. On one side vendors are pushing to maintain differentiation in their offers, on the other side customers are keen to get their uniqueness taken into account.

Applied to technology, we can consider that commoditization is reached when:

- The designs for implementation of the technology are widely available
- The technology has become easy to duplicate: competitors are all starting to build similar and compatible solutions
- Lots of alternative implementations of the technology compete directly
- The use of technology by itself does not offer a competitive advantage.

The Datacenter infrastructures already benefit from an important wave of commoditization in the 1990s, when most enterprises started to switch from large monolithic servers with proprietary architectures and Operating Systems (OS) to servers using standard (x86 architecture-based) and OS (Windows, Unix and Linux).

However, this switch from a “scale-up” to a “scale-out” approach (from upgrades through addition of components to upgrades through addition of new machines) resulted in a server sprawl, with a commoditized OS typically running a single application per server. These servers were typically under-utilized and demanded an important cumulated amount of power and cooling.

In the 2000s, this under-utilization rapidly drove customers to adopt, and vendors to propose, a “scale-in” approach based on blade enclosures (upgrades through addition of standardized nodes); optimizing power, space consumption and offering centralized management but also bringing back vendor lock-in.

Virtualization might appear as a solution to this lock-in. But even at the virtualization layer, swapping from one hypervisor to another is quite complex. Converting Virtual Machine (VM) formats is not that difficult, the issues come from detangling systems management software – backup, security, provisioning, capacity management and configuration management – that is deeply integrated into the hypervisor.

This point is crucial, because it is one the main driver for investing in Fabric-based Datacenters: solving system management segmentation issues, bringing consistency and uniformity from Datacenter components management (power, cooling...) up to application management.

The presence and strength of Open Source offers is usually a good indicator of a commoditization trend. Therefore despite resistance, commoditization of components is clearly on its way. An interesting movement is the emergence of open source concepts in hardware infrastructure with projects like Open Compute (http://opencompute.org/). Launched by Facebook in 2011 and backed by major players (HP, AMD, Salesforce.com, VMware ...), Open Compute aims to enable “the delivery of the most efficient server, storage and Datacenter hardware designs for scalable computing”.

Runbook Automation

Virtualization technology and especially the associated management tools provide the concept of software- defined services, where Datacenter components can be initiated and configured by software (virtual switches, firewalls etc.). This leads to the concept of sharing physical resources in a logically defined environment.

The adoption of Runbook Automation has pushed hardware vendors to provide Application Programmable Interfaces (APIs) that integrate with the provisioning layer and allow the combining of physical resources to a specific purpose at a given point in time, for a defined duration.

Today’s Runbook Automation stops where changes to hardware comes into play: The extension of a blade server by blades, the dedicated cable(s) between a computing unit and a storage unit as well as the setup of purpose-specific hardware, e. g. for a high performance (I/O) database cluster, are typical examples in a general purpose Datacenter environment where the use of Runbook Automation does not help as people must move the parts around.

The fully meshed Datacenter fabric truly disconnects physical upgrades from the way resources are provisioned. And Runbook Automation can cover the automated provisioning (or decommissioning) process end-to-end (see Figure 3).

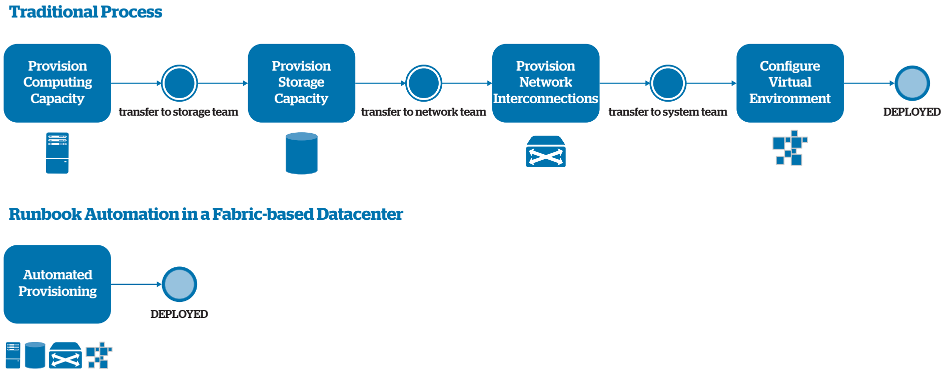
This promise has been around for a while and yes, has been implemented to a certain degree e.g. within a fully virtualized ecosystem but we recognize shortcomings are on a broader hypervisor scale and on Load balancer and Firewall configurations.

The Gain and Pain of using Vertically Integrated Stacks

Although most IT infrastructure investments are made on horizontal stacks with standardized interfaces to provide the largest compatibility with everything else beneath or above, we now see a strong trend to verticalization:

- Customers want to be serviced through domain or vertical business expertise rather than horizontal or technical expertise. They favor service and an SLA approach and in the context of economic uncertainty, they value offers that are mapped to their business models. What in any case happens on a macro-level such as the changing nature of outsourcing deals, concentrating more on

Figure 3: Provision Processes: Traditional vs. fully automated


Traditional Process
Provision Computing Capacity
transfer to storage team
Provision Storage Capacity
transfer to network team
Provision Network Interconnections
transfer to system team
Configure Virtual Environment
DEPLOYED
Runbook Automation in a Fabric-based Datacenter
Automated Provisioning
DEPLOYED

application level services than infrastructure beneath, hits the low level grounds of hardware design and Datacenter services

- Most industries face more and more complex environments in terms of business processes, regulations and competition. This increased level of complexity more and more requires dedicated IT solutions that take into account all the required specificities.

As a consequence, traditional software and hardware vendors extend their portfolio with "Vertically Integrated Stacks" (VIS) to be able to offer a top-to-bottom infrastructure and application architecture. This architecture aims both at better fitting to the needs of the customer and maintaining market shares in a space where large hardware and software investments are more and more driven by cloud and service providers who build multi-tenant platforms to host middle-market businesses.

As a result, verticalization drives the IT industry to move from:

- Horizontal marketing and go-to-market models toward more vertical, industry-focused models that identify business needs that are specific to individual industries
- A product-led sales model that focused almost extensively on technical sales to the customer's IT organization, to more of a services-led sales model in which more business-value-focused Account Managers engage as closely with the customers' business executives, as with their IT executives.

Do Vertically Integrated Stacks support Fabric-based Computing (FBC)? The company VCE (a close collaboration of VMware, Cisco and EMC) with their integrated vBlock concept has paved the way for 'Processing Area Networks'. They provide an integrated component (a 'vBlock') that is – from a resource and capability perspective – a Fabric-based Datacenter on its own. Its resources (I/O, CPUs, Memory and Storage) combine freely via software configuration.

But Vertically Integrated Stacks have two faces towards a Fabric-based Datacenter. Several hardware technology vendors focus on high performance appliance concepts; and applications are consequently specifically optimized to run on this appliance, an example is Weblogic on the Oracle Exalogic platform. Performance gains of this 'functionally engineered systems approach' are typically reached by bypassing abstraction layers imposed by virtualization technology. There are many other examples for application-specific VIS and – almost by definition – they create new 'resource pool islands' in a Datacenter with very limited chances to be ever merged into a General Purpose FBC design.

Convergence and the Evolution of Switching

Convergence is about the combination of computing power, networking and storage resources into aggregated pools with a global management platform. Until recently, however, most of these attempts were only partially successful due to a lack of integration between major Datacenter stacks. Cross-silo integration and orchestration is required through a common management platform in order to fully deliver on the business potential of Datacenter automation.

Convergence is addressing endemic problems in Datacenters today such as alack of high-level governance, idle workloads and resources wastage. In other words, it targets a radically simplified architecture that would offer seamless integration between loosely coupled storage, networking and servers by dis-aggregating resources into pools that are then reconciled with exact business needs.

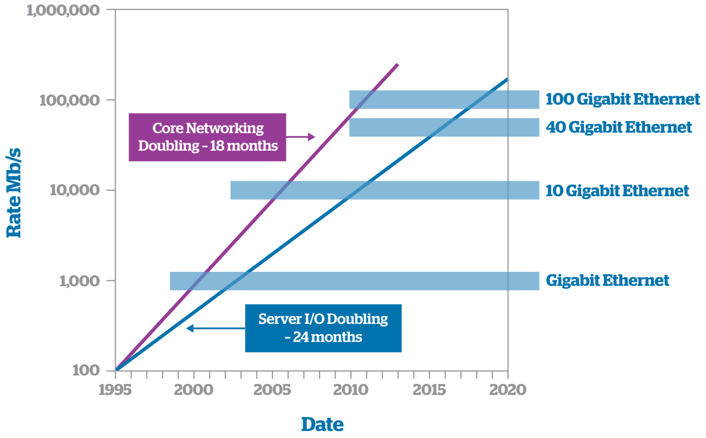
- Convergence of Computing Resources
- Convergence in Datacenter networking

Emerging evolution of switching

Even if today ROI benefits are still relatively unproven, a converged fabric should imply lower upfront capital costs, lower total cost of ownership, a more compact form factor and a greatly reduced amount of cabling. It should also drive to reduction in soft costs in areas such as infrastructure design, deployment and change management. Significant benefits will come from such an alignment into a single network.

- *The ability to embed existing technologies in a single cable*: FiberChannel, Ethernet and, in some cases, Infiniband could be supported onto a common set of physical elements.
- *Datacenter Bridging (DCB)*: Much like existing high performance fabrics, DCB infrastructures will connect multiple devices together into a fully "meshed" fabric with many possible physical paths.
- *40/100Gb Ethernet*: An advancement being driven by the rise in server and storage traffic on the converged network is the implementation of 40Gb and 100Gb Ethernet. This is the result of on-board 10GbE ports available to servers, so top-of-rack switches need 40Gb Ethernet uplinks or more in order not to become network bottlenecks.
- *Converged Network Adapters (CNA)*: Network interfaces that enable devices to connect to a converged fabric. Since CNAs are generally considered to simply address the merging of FiberChannel and Ethernet, some vendors use the term Unified Wire Adapter (UWA) to designate interfaces supporting full convergence.

Figure 4: 40 GE and 100 GE Computing and Networking [17]



The consolidation and interconnection of all the Datacenter equipment on a single converged fabric will be a big step. But beyond that, convergence will also affect the architecture of the components themselves, moving progressively from switching at I/O level to switching at system bus level.

Extending this low-level switching would open a large scope of possibilities, de-composing the traditional server architecture. CPU, memory, I/O bus and offload processors, such as graphics processors, could now become manageable into individual resource pools. Each of these pools could scale independently of the other, using standard interfaces for management, reporting and configuration.

Most Tier One hardware vendors are working on these concepts, but the emergence of a truly market disruptive solution is still constricted to a few technical challenges:

- Ultra Low latency: While switching at I/O level can support latencies measured in microseconds, switching at system bus level however requires nanosecond latencies, which put significant constraints on the maximum distance between nodes and the number of connections.
- Application-Specific-Integrated-Circuits (ASIC) performance: Evolution of ASIC technology will define a wide range of offload services integrated in adapters and host systems, from protocol acceleration to advanced adapters features. Getting highly performant ASICs is critical to the total infrastructure performance.

In-Memory Computing drives to extreme I/O

In-memory is a trend we see in a few popular databases, storing and handling data straight from memory in order to avoid any bottlenecks or latencies that disks or SSDs (Solid State Drives) might bring.

Many implementations are available today with different level of features. In order of complexity:

- Distributed Caching System – used to cache application data in a simple way. Implementations: Ehcache, Apache JCS
- In Memory Data Grid – distributed, partitioned and transactional cache. Implementations: Oracle Coherence, VMware Gemfire, JBoss Infinispan, Gigaspaces XAP
- In-Memory Data Database – fully fledged database working in memory. Implementations: SAP HANA, Oracle Exalytics.

Storage on disk or SSD is only used for database logging and backups to safeguard issues with the volatile RAM on power cuts.

In general with in-memory technology, the data is partitioned among the different servers to ensure horizontal scalability (with automatic back-up and recovery in case of node failure).

A high-end solution like SAP HANA requires hardware that is certified by the specific software vendor, where the software is optimized to use a specific technology and hardware configuration to optimize performance. This requires specific types of processors, memory and storage.

This shows that In-Memory Computing is very software-specific and based on specific hardware, having the advantage of really pushing the hardware to the edge of its capabilities.

If we would like to use the remainder of CPU capacity or even unused memory blocks for other applications in this context we have a need of high bandwidth and low latency networks to fulfill these tasks.

Currently networks speeds up to 100 Gbps are available, but are not yet up to the 500 to 640 Gbps we would need as a minimum to reduce latency of the memory over the network infrastructure, let alone the speed of the backbone to avoid bottlenecks there.

Also memory access from processors will improve and get faster in the future, Fabric-based Computing will need to bring solutions here and might need to break the traditional networks as we see in today's infrastructure, because these will become congestion points.

In-Memory computing will certainly challenge the Fabric-based Datacenter Infrastructures in flattening the infrastructure and flexible use of resources in the "Mesh or Grid" of resources. But might also bring a solution to the ever growing demands in flexibility, performance and availability of applications; even provide a flexible solution for the scale-out scenarios that are currently a challenge in the In-Memory arena.

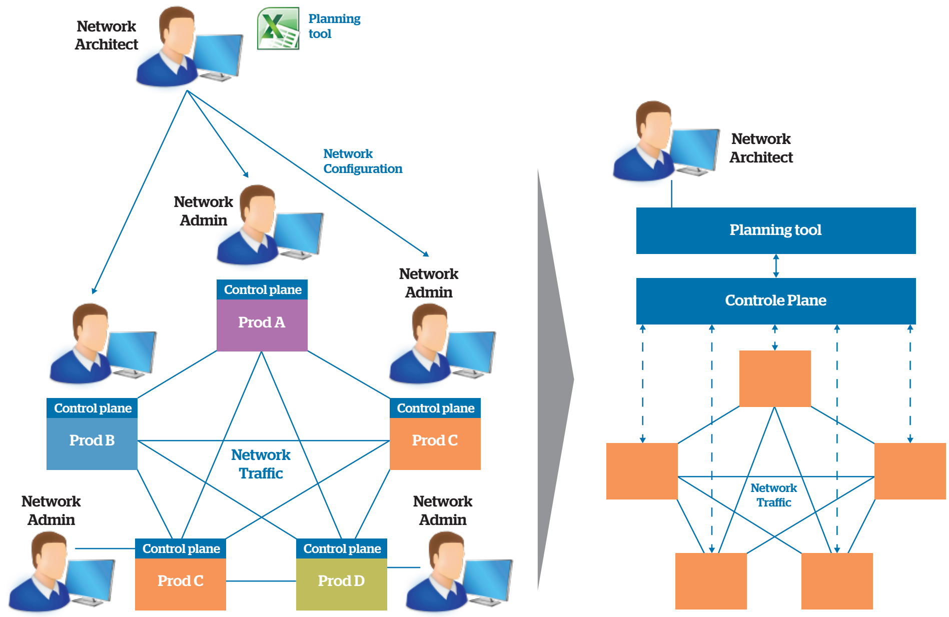
Software Defined Networking

The evolution of network switching in the Datacenter is taking place while another major trend in networking has emerged: software-defined networking. This trend takes its roots from the work done by several researchers in the OpenFlow project. With this project, researchers wanted to promote a way to perform large scale, real world experiments on alternative routing techniques while keeping the regular data traffic safe. It then materialized in the design of a method for switches to retrieve a set of routing policies from a central server when they encounter a data packet they are not used to dealing with. In a way, software-defined networking applies to IP networks a quite common idea in telecommunication networks: the split between the control plane and the data plane and the centralization of network control on a set of servers that manage routing at the network scale, pushing to the switches and routers the policies they should apply.

In the scope of a Datacenter network, using these concepts behind software-defined networking are tempting because in these networks, the link capabilities and characteristics are quite homogeneous, they are under the control of a single authority and are heavily monitored. Hence, the Datacenter Manager can use software-defined networking to set the routing policies in the network. This is even more important since the flattening of the switching architecture can have some drawbacks in a heavily connected environment if classical tree spanning policies are applied.

Many network equipment and virtualization solution providers have seen the potential of software-defined networking to enhance the way networks are managed in the Datacenter. 2011 and 2012 have seen many acquisitions of small startups by Cisco, Juniper or VMware to boost their activities in the software-defined networking space. This move tends to show that the general Fabric-based Computing movement reshuffles the cards between players that were previously complementary. Indeed, one of the major use cases for software-defined networking in the Datacenter is to ease the virtual machine duplication, migration or replication in the infrastructure, allowing the network to reconfigure more quickly

Figure 5: Operational Benefits of Software Defined Networking



than if a tree spanning re-computation or a routing algorithm procedure was run on the Datacenter's network elements.

Even if software-defined networking gained a lot of traction in the past months, this is clearly not a settled technology yet. Indeed, beyond the OpenFlow project, software-defined networking is not heavily standardized and the recent moves from companies tend to show that major networking equipment providers are still struggling to push their technological approaches. From a pure research perspective, a lot of open questions remain regarding software-defined networking. For instance, some research teams are working on the possibility of checking software routing code for reliability, consistency, and loop avoidance. As the programming logic can change quickly, it is important to have the proper tools to check that no 'poisonous' behavior is injected in the routing policies to apply in the network. Secondly, some researchers and early testers have expressed concerns regarding the delay in retrieving the routing policy from the central routing policy server used in software-defined networking systems. This latency might be problematic in heavily dynamic traffic conditions, as coming regular new flows might harm the ability of the network elements to retrieve the routing policy while switching the traffic properly. Indeed, software-defined networking appears to be quite efficient in relatively simple network switching reconfiguration use cases, while in more complex situations that can be encountered in Datacenters, involving VPN tunneling or firewalling, SDN still needs to prove their applicability, even if solutions with advanced functionalities like Lyatiss or Midokura MidoNet are emerging.

With those maturity issues in mind, Datacenter network designers might favor more mature alternatives applicable to a set of the use cases for software-defined networking, such as TRILL or MPLS-PCE that can help to solve some issues associated with routing traffic across two or more Datacenters efficiently, or apply aggressive update policies in the network elements responsible for service discovery.

Pairing nicely: Datacenter Infrastructure Mgt. and Fabric-based Datacenters

Fairly recently the idea of Datacenter Infrastructure Management (DCIM) is gaining additional momentum in managing Datacenters as an integrated environment. It changes the view of those responsible into looking at a Datacenter consisting of building-related technology that interacts with the IT components inside the building and vice versa. By creating this holistic view, DCIM will enable optimizations that are not feasible by looking at both domains in isolation.

Vendors of components and business analysts, as well as leading IT companies that manage large Datacenter facilities, are convinced that DCIM is a 'must have' for Datacenters. However, the majority of Datacenters are not yet using DCIM.

This contradiction is explained through our own and other research that reveals:

- That it is difficult to easily show a positive case for DCIM to individual cost and/or profit centers in a company, because the benefits are spread across different departments
- That implementation of a full DCIM system in such a way that the promised benefits are actually obtained, requires cross company collaboration, including facilities management, purchasing, IT management and external suppliers such as realtors and energy providers.

Typically we see deployment of DCIM of paramount importance to reach four levels of Datacenter management maturity:

- To create **integrated reporting** – how does the IT and underlying infrastructure (e.g. energy management, cooling, airflow, etc.) interact with each other?
- To **invest and support change** – by either human or automated analytics, can we derive results that support change in the infrastructure and explain why certain investments need to be made? This maturity level would also include asset and product life cycle management of all Datacenter components

- **For scenario-planning** – taking all the knowledge that exists in a DCIM environment, can we run simulations and predictions? And how can change affect our management objectives?
- **Enable Automation** – taking our scenarios and planning, based on existing data and information in our asset-management, can we now automate the proposed and earlier simulated change?

Recent investments in 2012 and 2013 by Siemens and Atos in collaboration have shown that we can indeed combine information from the lower Datacenter infrastructure with the data coming from managing the IT layers. By utilizing knowledge from, and experience in, complex event processing, large manufacturing systems and asset management in industrial plants and factories, we are able to build an integrated DCIM system that can support these new maturity levels.

We have observed short term results in lower energy costs, better energy usage and improved utilization of square meters, while maintaining the same power efficiency. We expect long term results to support the purchasing process, the change management in operational Datacenters and an increased availability through predictive maintenance.

DCIM will provide the capability to directly interact with the fabric through APIs and dynamically re-distribute workloads across the infrastructure, based on metrics received from the facility. This will allow for IT to be truly delivered as a utility, being able to match demand effectively and efficiently, without wasteful and expensive overprovisioning. Ultimately, combining Fabric-based Datacenters with DCIM will raise the maturity level of Datacenters, working towards the ultimate goal of creating fully dynamic, self-managing, self-optimizing, self-healing Datacenters.

Fabric on a Chip

Chips, which are "CPU only", have nowhere to go and no way to grow faster or be more powerful. They execute code on data pushed around at gigahertz speeds, both of which reside in memory shared between all the cores and from the perspective of an electron, very far away...;with more space eaten up by caches than by the compute cores modern 16 core chip, CPUs have hit another brick wall.

Today we stand on the brink of the industry rethinking its approach to growing compute power, especially since power consumption has become the major concern for even the largest IT companies like Microsoft, Google, Facebook and Amazon.

Two additional developments have helped to push the change:

High Performance Computing went to computing using graphic processors (GPU), where massive parallel chips sporting hundreds and thousands of reduced functionality cores easily outperformed the dozens of traditional CPUs you could buy for the same amount of transistor real-estate and the same amount of electrical power. These GPU designs reduced clock rates to around 1GHz and upped the number of compute nodes by eliminating most of the real state used for cache on 4GHz traditional CPU designs.

Smartphones are now matching the performance characteristics of a 10 year old workstation, but use energy in the tiny amounts possible by the form factor and battery power. It was enabled by "dark silicon", a technology which doesn't attempt to squeeze out the last drop of compute performance by running the chips as fast and hot as possible with the most advanced cooling techniques, but which aims to shut off every single transistor which doesn't contribute right now, while still making billions of transistors available for millisecond peaks.

Emerging CPU designs combine and enlarge Systems on Chip (SoC), GPU compute and dark silicon ideas, enlarge NUMA (Non-Uniform Memory Access) to actually tile or partition DRAM between CPU cores and use the HPC type fabrics to map and reduce compute workloads: On a chip.

When you are running an Internet-facing enterprise application, your setup will most likely include a whole cascade of North-South facing switches, appliances and servers for firewalling, load-balancing, proxying, application and database servers. Today these are using high-performance x86 CPUs running at multiple GHz and then move data leisurely through copper Ethernet cables across the Datacenter. With a Datacenter- on-Chip (DoC) the entire cascade runs on extremely power efficient single GHz ARM CPUs, but every hop in the cascade sits on the same chip and is only a few nanoseconds away.

The big difference is power consumption and therefore cost: While the x86 CPUs were designed for compute throughput, are shared between as many clients as possible and therefore run 'hot and thirsty', the Datacenter-on-Chip is designed for latency, tiny enough to almost give each client its own and 'cold' unless used. In a DoC setup dark silicon and next generation power rails may eliminate the need to move the logical systems, while smart fabrics are still needed to direct the traffic in the most energy efficient way while ensuring response times.

This development is driven by vendors currently selling network silicon. Today these are just another brand of Systems on Chip where most of the chip area is dedicated to massive parallel programmable fabrics capable of routing terabits/sec at wire speed to dozens of ports with just the latencies required to decode the incoming packet headers and some general purpose CPUs for management. They are designed to cascade and mesh to form fabrics with thousands of ports inside big core switches. Tomorrows network silicon vendors just add lots of 64-bit ARM v8 CPU cores on the die space made available with the next process shrink.

On these fabrics, PCI peripheral busses, Ethernet links or even NUMA memory access become just protocols and the logical hardware topology is defined in software while the physical hardware topology is on the chip.

Some examples:

Tilera GX packs up to 72 CPU cores, 8 10Gbit XAUI and 32 1Gbit SGMII Ethernet links, 24 PCIe lanes and four channels of DDR3 1866 on a single chip. Full and independent SMP systems can be created combining any number of these elements into pools or tiles, each running their own OS instance and implementing a typical full vertical enterprise web application stack on a single chip with plenty of bandwidth for extending the architecture horizontally. On chip IP is capable of terabit wire speed cryptography and packet reassembly.

Calxeda Energy Core – Calxeda is using four ARM A9 32-bit CPUs per chip, includes a full 128-bit floating point as well as eight 10Gbit Ethernet links, 16 PCIe lanes, 5 SATA ports. Each chip includes a fabric, which can be configured at boot time to link these chips into SMP, coherent or shared-nothing (network only) designs. Facebook is replacing its x86 servers with these: The complete application tier.

Cavium Project Thunder – Cavium is a network silicon vendor offering solutions for the highest speed Ethernet and LTE networks witches including line-rate security IP-blocks. As network silicon they are designed to stack to thousands of ports at minimal latencies and maximum bandwidth. In the past they used ARM and MIPS cores on these network SoCs only for management functions. Recently they launched Project Thunder, which is about adding compute to their network silicon using the next generation 64-Bit ARM v8 cores.

Unlike GPUs the systems on these chips are capable of running ordinary Linux software. While the Tilera is still optimized towards certain network use cases, for Facebook the main reason to choose Calxeda is the ability to run its applications unchanged. ARM's 64-Bit v8 CPU is capable of running any of today's enterprise applications as well as the most demanding emerging in-memory computing stacks.

Processors have entered the way towards general purpose computing, integrating the full fabric on chips.

Business Impact

Increased business needs and a faster change of business speed, will add pressure to the underlying infrastructure of any IT used by the business.

Businesses need more and more compute power, and they need it at the right time to generate a competitive advantage. IT departments will implement Fabric-based Datacenters to cope with these demands:

- Greater agility and shorter time to execution
- Cost reduction through better resource utilization and automation
- ”Right sizing”, allowing previously achievable levels of flexibility.

Solving these challenges will enable faster alignment of IT resources to changing workload patterns, time zones, business priorities and any other changing demands that may arise.

Fabric-based Datacenters will bring about improvements in performance and optimization that are not possible by addressing only selected parts of the infrastructure stack, even though advances with in-memory computing are bringing clear benefits for selected processing needs. Considering the demands(sometimes near-real-time) that will come with topics like Big Data, Pattern Based Strategies and Contextual Smart Mobility, the pressures to develop viable Datacenter Fabric solutions will be significant even if in the meantime we see it being developed alongside with specific architectures that address performance or point issues.

Market Adoption

Vertically integrated stack implementations are reaching a point of maturity within the market and if universally deployed, could offer a common architecture to build a complete Datacenter Fabric.

However, this may come at the cost of a technology lock-in, preventing seamless integration of standard components or scalability beyond “forklift” upgrades if markets do not force the setup of interoperability standards as we have been seeing in other technology trends before, like the rise of X86 architectures… The major database vendor shave also paved the way to network convergence and fabrics, but they have defined application sets baked in, so they come with inherent limitations of having a specific purpose. The design of the true Datacenter Fabric will therefore need careful consideration to fully achieve the multipurpose flexibility and agility it promises.

Challenges

In order to enable Fabric-based Datacenters as a ‘de facto’ approach for datacenters, a number of challenges will have to be overcome:

Vendor lock-in is probably the most challenging one. Current solutions are largely proprietary, with limited hardware interoperability and customized management interfaces. If standards do not emerge, Fabrics might become the new mainframes, with vertical silos and a significant cost of change.

Another key point will be the *availability of operating systems and hypervisors that are able to fully leverage the scalability* and the granularity of Fabric-based Datacenters. Current mainstream operating systems are not yet ready to fully exploit the benefits of fabrics. By extension, availability of ‘fabric-aware’ application development frameworks will be an important step to quickly build applications that could leverage a true fabric.

With Fabric-based Datacenters, the architecture becomes “software” and is then subject to multiple security threats. Security is still one of the top drivers to maintain physical segmentation and isolation between platforms – in several cases even required by industry security standards such as PCI.

Fabric-based Datacenters are by definition dynamic and shared, which cannot be addressed with traditional security methods. At data level, data protection will be a challenge, both from a technical and legal standpoint.

Proper transition and change management will be essential. Fabric-based Datacenters will require important organizational changes, breaking down the borders between development and operations, between software and hardware, between network, storage, server and management layers.

Conclusion

The concept of a Datacenter Fabric, takes optimization and flexibility to the next level, extending the orchestration of the complete IT scope to include the complete Datacenter. Fabric-based Computing is not a reality yet but the concept and a number of the building blocks are firmly established. It promises to address the goals of producers and consumers of IT services alike, with a 'managed by software' approach, based on continuously (re)negotiated terms between providers and consumers.

Still, having very large highly interconnected environments, managed by software raises major challenges in terms of reliability and resiliency. It is critical to avoid those software bugs or hardware failures which create major incidents with cascading effects due to the lack physical segmentation.

Whilst there is still more to do with APIs, security concepts and completion of all the necessary levels of integration, the ability to deliver the long promised full stack of "utility" and "on-demand" computing is on the horizon and the vision of having a real Foundation IT that can give businesses the power and agility they really need to fulfill their business strategy with the proper IT systems is coming closer and closer.

References

[1] Robert Marinus van Wessel *“Realizing Business Benefits from Company IT Standardization”* (2008)

[2] DigitalRealityTrust *“What is driving the US Market”* (2011)

[3] Cisco *“Cisco UCS: A Real-World TCO Analysis”* (Apr 2011)

[4] Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz André Barroso *“Failure Trends in a Large Disk Drive Population”* – Google (2007)

[5] Ole Hanseth and Kristin Braa *“Hunting for the Treasure at the End of the Rainbow: Standardizing Corporate IT Infrastructure”* (2009)

[6] Sven Graupner, Jim Pruyne and Sharad Singhal *“Making the Utility Datacenter A Power Stationfor the Enterprise Grid”* – HP Labs (2003)

[7] Rolf Enzler, Christian Plessl, and Marco Platzner “Virtualizing Hardwarewith Multi-Context Reconfigurable Arrays” (2004)

[8] Morgan Stanley *“Blue Paper: Cloud Computing takes off”* Morgan Stanley Research (May 2011)

[9] Albert Greenberg, Parantap Lahiri, David A. Maltz, Parveen Patel, Sudipta Sengupta *“Towards a Next Generation Datacenter Architecture: Scalability and Commoditization”* Microsoft Research (2008)

[10] M. Kamoshida, T. Hayashi, S. Takagi *“Next Generation Datacenter Outsourcing Services”* FTS (2010)

[11] Ronald Luijten, Cyriel Minkenberg *“Viable opto-electronic HPC interconnect fabrics”* (2005)

[12] M. Lin, J. Hsieh, D. HC Du *“Performance of High-Speed Network I/O Subsystems: Case Study of A Fibre Channel Network”* (1994)

[13] Mark White, Bill Briggs *“Tech Trends 2011 – The naturalconvergenceof business and IT”* Deloitte (2011)

[14] J. Young, S. Yalamanchili, *“Commodity Converged Fabrics for Global Address Spaces in Accelerator Clouds”* The 14th IEEE Conference on HPCC (June 2012)

[15] Cheng-Zhong Xu, Jia Rao, Xiangping Bu *“A Unified Reinforcement Learning Approach for Autonomic Cloud Management”*, Wayne State University, 2012

[16] DCIM contribution in Section 4.8 credit to Mark Hensbergen, Senior Architect, Atos NL.

[17] Gautam Chanda *“The Market Need for 40 Gigabit Ethernet”*, Cisco Public Information (2012)

About Atos

Atos SE (Societas europaea) is an international information technology services company with annual 2012 revenue of EUR 8.8 billion and 77,000 employees in 47 countries. Serving a global client base, it delivers IT services in 3 domains, Consulting & Technology Services, Systems Integration and Managed Services & BPO, and transactional services through Worldline. With its deep technology expertise and industry knowledge, it works with clients across the following market sectors: Manufacturing, Retail & Services; Public sector, Healthcare & Transports; Financial Services; Telecoms, Media & Technology; Energy & Utilities.

Atos is focused on business technology that powers progress and helps organizations to create their firm of the future. It is the Worldwide Information Technology Partner for the Olympic and Paralympic Games and is quoted on the NYSE Euronext Paris market. Atos operates under the brands Atos, Atos Consulting & Technology Services, Worldline and Atos Worldgrid.

atos.net