THE SELF-AWARE HYBRID DATA CENTRE

The first step to knowledge is to know we are ignorant.

Written by Chris Mellor.

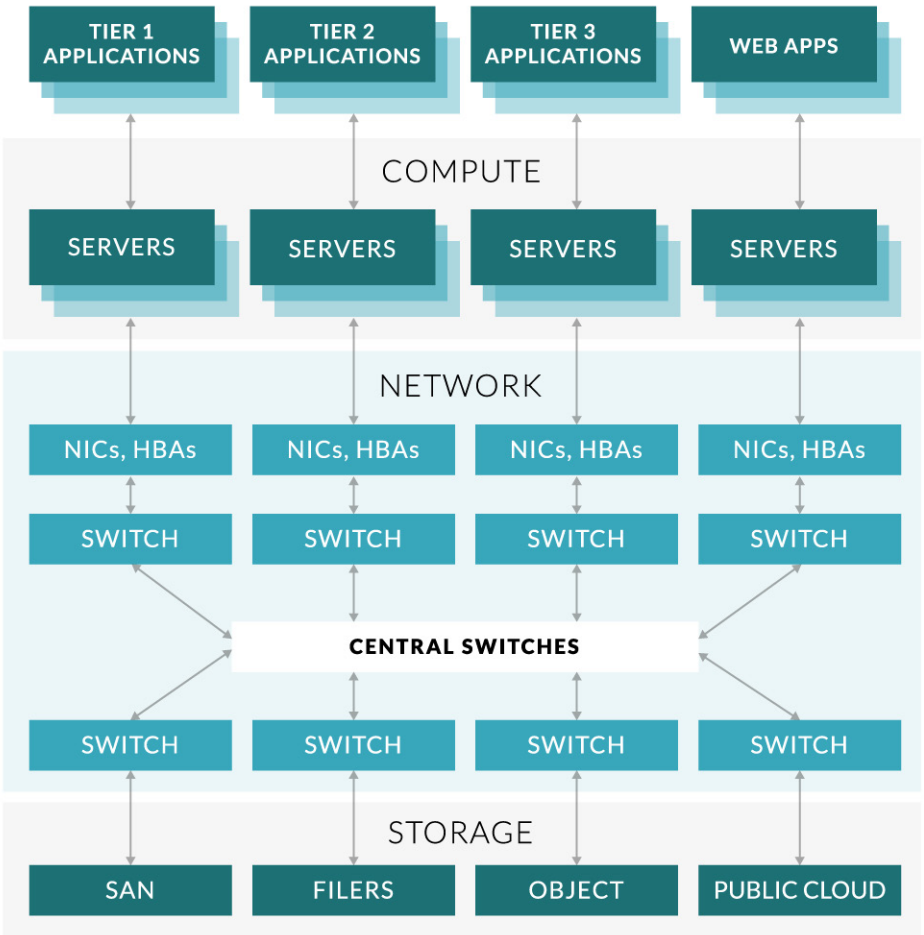
virtana

The Register®

The autonomous, app-aware, self-managing and self-healing data centre is here. Data centres are incredibly complex systems which need far more penetrating, detailed and fast-reacting real-time management than humans can provide. Artificial Intelligence (AI) working on data from every level of a data centre's application, server, networking and storage infrastructure, is the only way to realise this goal and bring proper business management rigour to the operation of these mission-critical systems.

The purpose of a data centre is to run an organisation's applications efficiently and effectively. Any problems in the increasingly hybrid data centre infrastructure preventing this need to be detected and rooted out.

An organisation's data centre is quite possibly the most complicated machine it operates in terms of the overall breadth, depth and connectivity of its many components. A business with several data centres and using several public clouds is operating one of the most complex IT systems on the planet. There can be tens of thousands of application services. The number of individual events occurring in a hybrid, multi-cloud data centre estate every day, from application events down to packet-level operations on network cables is in the trillions per day, possibly even greater.



The hybrid data centre is a layered system.

A data centre's operation should help the business run the business. Yet many companies are not sure how data centres should be run because the managers and operators don't know what's going on inside them. That which cannot be monitored and measured cannot be managed effectively.

CHALLENGES OF MANAGING THE HYBRID DATA CENTRE

Managers and operators of hybrid data centres may have only a partial grasp of what is happening inside a data centre for several reasons:

- Few data centre component layer hardware and software devices tell them how they are interdependent, interconnected, and interrelated to other components.
- There is no clarity on how applications use the infrastructure components they need, and how infrastructure component problems affect application performance.
- Applications are not ranked by business priority or service objectives into, for example, gold, silver and bronze tiers.
- There is no way of understanding the relative importance of dealing with infrastructure component problems in terms of handling the most important applications first.
- Legacy tools have little ability to predict when problems will occur.

It makes no sense to pour resources into fixing an infrastructure component issue affecting bronze tier applications when a gold tier application requires attention.

Picture credit: Constant Loubier.

Without the ability to tier applications by importance and understand their infrastructure footprint and infrastructure component use and dependencies, it is impossible to manage a data centre effectively. It means no optimising for performance, efficiency and cost. It also means that understanding the growth in demand for infrastructure resources will be inadequate and render resource growth planning ineffective.

It is impossible to manage a data centre effectively without being able to tier applications by importance and understand their infrastructure footprint and infrastructure component use and dependencies. Without this, there can be no optimising for performance, efficiency and cost and a lack of understanding of the growth in demand for infrastructure resources is likely to severely hamper growth planning.

This issue is exacerbated (2X) when private data centres are interconnected with each other, exacerbated again (3X) when linked to a public cloud supplier, and yet again (4X) when multiple public cloud suppliers are involved.

Unless business IT executives have access to clear and accurate assessments of data centre operational costs and trends, they will be unable to effectively optimise and manage the deployment of applications to data centres within their hybrid and multi-cloud IT environment.

That means that their hybrid and multi-cloud data centre infrastructure must present operational metrics and statistics to a management system that can implement policies set by executives and that can then report on the degree to which these objectives are being met.

OPERATING BLIND

A company's finance and accounts function is a marvellous precision machine whose operations and procedures have been honed and perfected by decades of experience. The whole point of this experience is that financial events and decisions should be visible, and their interdependencies known and understood. Where money flows, intelligence goes and all relevant stakeholders are alerted to what is going on, how budgets are affected and how the overall financial picture is developing.

It's different in most data centres. Often management intelligence does not know where bits and bytes flow, or how fast, or why, and, worse than that, cannot find out. Where data flows through the IT stack, intelligence does not always follow. Real-time visibility is essential. Tools such as log analysers are insufficient as they are reactive by definition.

Unlike a power station or a car manufacturing plant built for one thing, a data centre is a general-purpose infrastructure machine, built to run myriad application services, some decades old, others quite new, with separate layers allowing supplier choices, but also providing dividing walls between software-vamped hardware devices. When a component fails the consequences can be devastating.

OUTAGE AND INEFICIENCY COSTS

A business can suffer huge losses if there is a mission-critical application outage. British Airways faced an £80m cost from a May 2017 IT outage that caused 726 flights to be cancelled over three days and resulted in 75,000 passengers stranded in the wrong places.

When an outage hit Delta Airlines in 2017 it cancelled 280 flights and its losses reached $150m.

A 2014 Gartner analysis said IT outages could cost $5,600 per minute, and up to $300,000 per hour in web application downtime. These numbers will be larger now.

Prevention, therefore, is key. Without forward looking strategic monitoring and analytics platforms deployed in the infrastructure, outages will happen. The overarching need is to be able to diagnose and fix them quickly – before the business is impacted. As businesses become digitally transformed, an ever-greater proportion of their operations depend on the IT business applications running on the computers, storage, networking and cloud infrastructure.

As outages have the potential, in extreme cases, to paralyse a business, alongside the crucial diagnose and fix processes there must also be IT disaster recovery provisions.

THE WAR ROOM SYNDROME

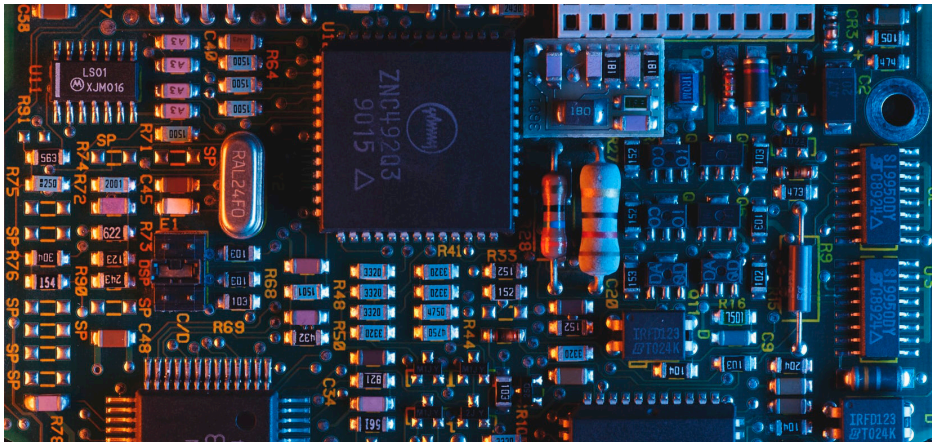
When a serious application performance problem arises, such as a mission-critical outage, and the cause is not immediately clear then an IT problem war-room may be set up, with representatives from each of the IT functions involved, such as networking, storage, servers and applications.

However, with heterogeneous suppliers' kit each having its own discrete diagnostics and reporting function, these gatherings can degenerate into defensive finger-pointing instead of focusing on finding the root cause of the problem.

The fundamental issue is that there is no neutral, independent tracing of application data flows through, and related events in, the IT stack. It's a cliché but the staff responsible for support and diagnostics could be operating blind. They need sight and are not getting it.

INFRASTRUCTURE COMPLEXITY

Let's suppose there is a problem with the third 16Gbit/sec port on a network switch and that is slowing down network packets of a particular type. These happen to be originated by a particular server but only when it is running a virtual machine executing an application which ensures customer orders are dispatched from a warehouse.



Picture credit: Umberto.

The switch port-server-VM problem causes the order dispatch rate to slow down but without an understanding of the data flow route through the stack, and the related event logs at each stage, the war-room is looking for needles in haystacks in the dark.

The whole stack needs instrumenting from the point of view of measuring an application's performance as its data flows pass through the stack.

It needs to know that, when that application VM runs on that server, its data flows will take particular tracks through the infrastructure's layers and devices within each layer. The application's service level agreements (SLA) need to be known so that, if its performance departs from those SLAs then the issue can be detected quickly and diagnostic routines started.

These would look at event logs and device states for that specific application's track through the infrastructure layers, scanning for any inconsistencies and issues that relate to the problem. If, for example, port 4 of the network switch is normally used but is congested or down and port 3 is used instead, then this needs to be known, so that that port 3 event logs can be inspected straight away.

War-rooms are essentially the result of failure, put in place after an IT disaster. The only realistic option for preventing disasters happening in the first place is to use an intelligent, application-aware and real-time monitoring system - Artificial Intelligence for IT Operations (AIOps).

INSTRUMENTATION

Instrumentation on its own is not enough. It can tell us what's going on at any moment in the many layers of the data centre stack, the deep infrastructure, and the components within them - the data centre's topology. But this just generates torrents of data values for component devices and states.

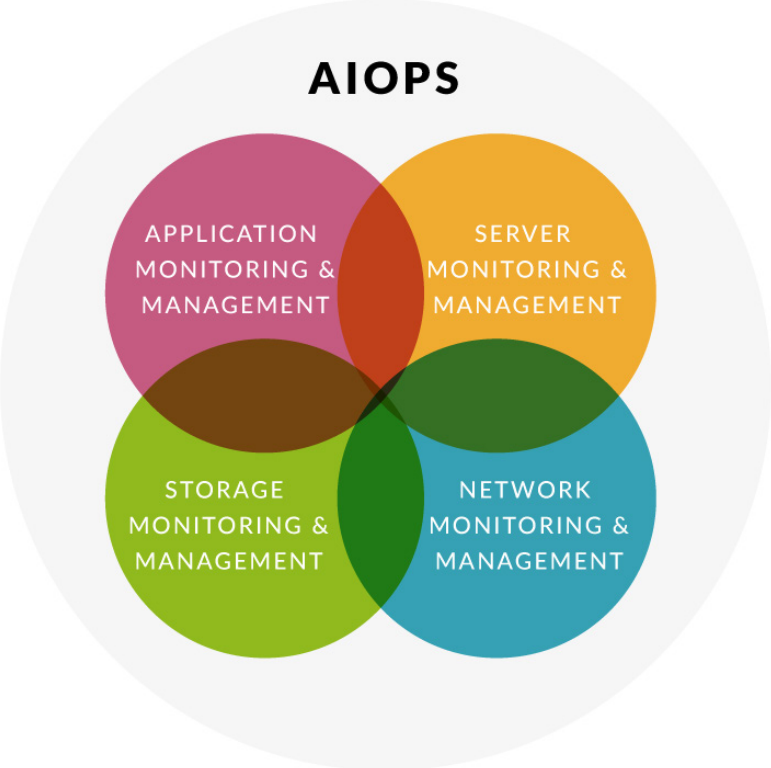
Without understanding the relationships and dependencies between these components and layers and the applications they serve, we cannot see how faults and behaviours in one part of the stack can ripple through it and affect application performance in particular ways.

Four things prevent us from doing this: the sheer amount of information; its arrival rate; its complexity; and a lack of understanding of the application-infrastructure topology.

Effective human control of a trillion-events-per-day-level hybrid data centre infrastructure deployment is clearly not possible. The best way to deal with vast amounts of complex data arriving at a torrential rate through particular paths is to use computer power, and that's where AI and Machine Learning (ML) come in.

AIOPs

Management consultant and thinker Peter Drucker in 1967 said: "The computer makes no decisions; it only carries out orders. It's a total moron." Picture Caption: AIOPs diagDrucker was right in 1967 then but is wrong now. That is because ML enables computers to expand beyond simple "if X then Y" procedures and learn from what is presented to them. That means they can decide what is the meaning of a vast set of complex and changing variables and either respond automatically or present choices to human operators.



AIOps distilled.

AIOps is a new approach, but limited in its usefulness unless it is married with a real-time monitoring system. It's a truism, but better data really does give better insight, leading to better results. First generation AIOps products (for example, Moogsoft, Big Panda) are reactive and can't be used for proactive performance and capacity management.

ML is blindfolded without an understanding of the application-infrastructure topology and the relative importance of the many, many applications in a data centre. Once it has this understanding, it can then be the single source of truth for infrastructure events.

When this knowledge is applied to the torrent of deep-stack events in a data centre, ML can quickly detect any inappropriate event pattern, help root out the infrastructure cause for application slowness and help deal with it.

Cars, vans and trucks all now have Digital Motor Electronics (DME), which instrument the many systems in the vehicle - engine, exhaust, airbags, and so forth - and logs event readings. Engineers can plug in DME reading systems to diagnose and fix vehicle problems.

AIOps is the data centre DME equivalent, augmented by AI and capable of diagnosing, fixing or mitigating, thus alerting us to problems in real time. It can detect resources running out of capacity in the data centre's stack layers, or devices beginning to misbehave, and pre-empt their failure.

VIRTANA

Virtana's roots are in infrastructure performance management, with real-time analytics. The company understands that enterprises, becoming more and more dependent on IT, need a single platform to monitor, manage and resolve data centre issues.

For Virtana, AIOps is a key capability of its hybrid infrastructure management platform. Virtana combines real-time infrastructure monitoring with AI-powered analytics to help its customer quickly resolve problems, forecast infrastructure capacity needs, and automatically balance workloads to infrastructure resources.

That platform has been designed to supply an independent perspective and to overcome the inherent limitations of server operating system and hardware, virtualisation, storage and networking suppliers.

Only a third party with a combination of bottom-up instrumentation for deep infrastructure visibility, simulation expertise, workload performance analysis, application awareness and AI/ML capabilities, can provide what's needed. That is the full stack instrumentation, topology mapping, monitoring, management, pre-emptive alerting, diagnostic and fix capabilities needed by a modern data centre operating in a hybrid, multi-cloud world.

Virtana is such a vendor and has a clear vision of what is needed and how to develop and supply it. It knows that the AIOps platform has to be designed for mission-critical environments at enterprise scale and provide a combination of monitoring, capacity and cost management, and workload automation in one product line.

THE FLUID ENTERPRISE AND AIOPS

Enterprises are fluid and will often move workloads as needed between the on-premises private cloud and more public clouds. The AIOps platform has to operate knowing this is the case and be able to reach into each one of these environments. The end-point is to bring a degree of self-awareness, of autonomy, to a business's interconnected public and private IT facilities so that they operate much more smoothly. The administration and the management of hybrid data centres needs to be digitally transformed and the Virtana infrastructure monitoring and AIOps platform is the only feasible way to do it.