FREEFORM DYNAMICS

Inside Track Research Note

In association with

FUJITSU

Ceph as an Enabler of Growth and Scalability

Hyper-scale should not mean hyper-cost and complexity

July 2015

In a nutshell

A perennial problem with storage is how to deal with escalating requirements in a smooth, manageable and non-disruptive manner. By removing many of the traditional limits on system expansion, Ceph based configurations allow your storage to scale painlessly and inexpensively. You can better keep pace with new and changing requirements, without the fear of running into that next 'wall'.

About this Inside Track

The insights presented in this document are derived from independent research conducted by Freeform Dynamics. Inputs into this include in-depth discussions with IT vendors and service providers on the latest technology developments, along with intelligence gathered from mainstream enterprises during broader market studies.

The need for smoother capacity growth

Data growth is not a new phenomenon. Organisations have been storing more digital information year-on-year since the very early days of computing.

IT teams have traditionally kept up with this through periodic upgrades or extensions to their storage infrastructure, a process with which you will undoubtedly be familiar. Every year or two you add more capacity to core storage systems, aiming to incorporate enough headroom for the next couple of years. This cycle repeats until you eventually reach the expansion limits of your existing setup. By that time, the hardware is usually quite old, so you invest in a bigger system, migrate your data, and the growth management game starts all over again.

The trouble is that accelerating growth rates have progressively shortened the time between expansion and upgrade events. As each event brings significant cost, distraction and risk, this traditional lurch-by-lurch approach to keeping up with increasing demand is becoming less sustainable. A new system you put in place today might easily become maxed-out from an expansion perspective well before its natural end-of-life, unless you purchase a configuration way beyond your current needs.

A new system you put in place today might easily become maxed-out from an expansion perspective well before its natural end-of-life.

In response to this, storage vendors have been working to make familiar storage array technologies more scalable and expandable. Coupled with advances in virtualisation and other techniques, highly flexible pooling and allocation of storage resources is now possible. This not only allows you to manage growth more effectively, it also enables the rapid provisioning of capacity on demand, a requirement that's becoming increasingly important. As a result of digital business transformation and the adoption of Agile and DevOps methodologies, today's application landscape is subject to a level of continuous change and expansion never seen before.

Against the backdrop of this more challenging and dynamic world, however, the evolution of traditional storage offerings is just one part of the story. Whole new architectures conceived from the outset to deliver high degrees of flexibility and responsiveness have emerged. Based on scale out principles, these handle growth a lot more elegantly and cost effectively. They also directly tackle the common performance, management and data protection requirements that arise as storage volumes grow into the Petabyte range and beyond. An example based on this kind of scale-out approach is Ceph.

New architectures conceived from the outset to deliver high degrees of flexibility and responsiveness have emerged.

Introducing Ceph

Ceph can be summed up as an open source software solution conceived and built to deliver hyper-scale performance and almost limitless expandability. If you are working in a telco, media or cloud service provider environment, such words are probably music to your ears. If you are responsible for storage in a more 'mainstream'

Ceph can be summed up as an open source software solution conceived and built to deliver hyper-scale performance and almost limitless expandability.

enterprise context, however, the words 'risk' and 'overkill' may spring to mind. But please read on. Deployed in the right way a Ceph-based solution could be just what you need to break out of the constraints you have been living with, and it needn't be costly, risky or time-consuming to implement - quite the opposite, in fact.

We'll come onto the practical implementation considerations that make this possible shortly. Before we get into that, however, let's take some time to understand the basics of what Ceph is, how it differs from the traditional storage solutions you are used to, and why this might be relevant to your business.

The Ceph architecture achieves exceptionally high throughput by parallelising reads and writes, and very high resilience by combining 'replicate on write' functionality with 'self-healing'.

A conceptual architecture view

At a conceptual level, Ceph is pretty easy to understand. Starting at the bottom of the stack, each installation is underpinned by a set of clustered resources that powers a distributed object store capable of handling any type of unstructured data. The Ceph architecture achieves exceptionally high throughput by parallelising reads and writes, and very high resilience by combining 'replicate on write' functionality with 'self-healing' capability.

Ceph is not just about object storage.

Sitting on top of the storage cluster is an object gateway that publishes Swift and S3 compatible APIs. But Ceph is not just about object storage as some may assume. Alongside the object gateway is a block interface for hosting and accessing virtual disks, along with a distributed POSIX style file system that exploits the parallel processing capability of the underlying cluster (Figure 1).

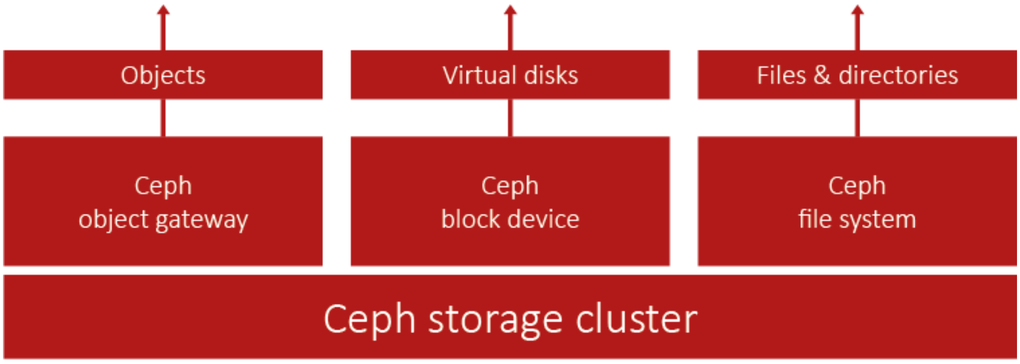

Figure 1
A conceptual architecture view of Ceph

Like all concept diagrams, this one is no exception, in that it doesn't convey what goes on behind the scenes. Let's therefore explore some specifics.

Ceph itself is purely software.

Very clever 'software defined storage'

Ceph itself is purely software. It needs to be brought together with the necessary hardware to do anything useful. This can be based on any disk and (x86) server technology that conforms to common industry standards. The end result is an open 'software defined storage' (SDS) environment. But Ceph isn't just a software replacement for a traditional hardware based storage controller. It goes much further than that.

Ceph isn't just a software replacement for a traditional hardware based storage controller. It goes much further.

If you take a close look at a Ceph cluster, you'll see it is made up of a number of nodes, each of which brings together a unit of CPU with a unit of disk. Combining compute and storage in each building block is a key enabler of Ceph's parallel processing capability, and the way this works is very clever.

The majority of the nodes in a Ceph environment take the form of Object Storage Daemons, or 'OSDs', which each store and manage a fragment of the data held on the cluster. When an application needs to store an object, OSDs 'collaborate' to work out how the primary data (and typically two secondary copies) should be distributed around the cluster, the idea being to keep everything balanced and optimised. When an object read is requested, OSDs similarly collaborate to retrieve the necessary data fragments (in parallel) and serve them up appropriately.

Data placement, retrieval and recovery is coordinated in a peer-to-peer manner across tens of thousands of nodes, with no separate controllers, and thus no bottlenecks or single points of failure.

The secret sauce here (or not so secret given that Ceph is open source) is an algorithm known as 'CRUSH'. It's beyond the scope of this article (and its author!) to explain how it works in detail, but it's what it enables that matters. Data placement, retrieval and recovery is coordinated in a peer-to-peer manner across tens of thousands of OSDs, with no separate controllers, and thus no bottlenecks or single points of failure.

This core capability is exploited within the higher level components that go to make up the Ceph environment. Block storage can be thin provisioned, for example, which enables rapid cloning of virtual machines in a hosted server context, which is great for service providers, but also very useful for enterprises looking to consolidate virtual server images. The spin off benefit of centralising in this way, apart from cost saving, is the flexibility to migrate VMs between physical hosts with ease – even 'live migrate' if the hypervisor supports it.

One of the great benefits of Ceph is its hardware-related flexibility.

Another example of a cool Ceph trick is something known as 'dynamic sub-tree partitioning' which is used to continuously optimise the distributed file system. Again based on the principle of peer-to-peer collaboration, a series of metadata nodes work together to share the load in a balanced and optimised manner as file systems grow and evolve – with parallel processing performance and no single point of failure.

Smoothing the growth and scalability curves

Coming back to the topic of growth and scalability, one of the great benefits of Ceph is its hardware-related flexibility. The compute elements in a cluster can be high-end or low-end CPUs, while the storage components can be spinning disks (JBODs), flash drives or even existing arrays. The choice of hardware is driven by factors such as performance requirements, maintenance objectives and cost constraints. Having said this, as a fully fault-tolerant, self-healing, parallel processing environment, Ceph will deliver resilience and performance on even the most basic of commodity equipment (though the use of cheap generic components may often represent false economy from a maintenance overhead perspective).

Any number of new resources can be added to the cluster at any point in time.

In terms of growth, Ceph can scale linearly from tens, to tens of thousands of potentially heterogeneous nodes in a cluster. Any number of new resources can be added to the cluster at any point in time, e.g. the steady addition of a few nodes at a time to stay ahead of a smooth growth curve, or a big chunk of aggregate processor power and disk capacity to handle an exceptional need for expansion. Whatever is added will be absorbed and made use of immediately and automatically. A small number of 'Monitor' nodes collaborate to keep track of any OSDs added to the cluster, or removed, either deliberately or as the result of a failure. Monitors also maintain a real-time view of the cluster state, e.g. which nodes are up or temporarily down.

A Ceph installation can grow and scale at whatever pace suits your business, minimising the need for incremental cost, re-engineering and disruption to production systems.

The upshot of all this is that a Ceph installation can grow and scale at whatever pace suits your business, minimising the need for incremental cost, re-engineering and disruption to production systems. And, of course, there are no inherent limits to force you into an expensive and disruptive forklift upgrade and migration exercise.

Too good to be true?

We have covered some of the key concepts behind Ceph not because you need to know the mechanics to make use of it, but to provide some insight into the automation features that can make life so much easier for you. Ceph was designed with ‘lights out’ operation in a high end service provider environment in mind, but the same near ‘black-box’ simplicity is also great for even smaller scale requirements.

The benefits of Ceph are dependent on having a properly configured installation in place, and constructing one of these from raw components is not for the faint-hearted.

But the benefits of Ceph are dependent on having a properly configured installation in place, and constructing one of these from raw components is not for the faint-hearted. You need the right skills and resources, and a willingness to interact with a very lively, but constantly evolving open source project.

The good news is that some of the more forward-thinking mainstream storage suppliers are delivering preconfigured solutions.

The good news is that some of the more forward-thinking mainstream storage suppliers are delivering preconfigured solutions that bring the Ceph software together with the right processor, storage and networking hardware so you can get up and running quickly on a fully supported basis. These give you all the advantages of Ceph, without the disadvantages of the ‘Do it yourself’ (DIY) approach. We expect such solutions to become more widely available over time, but for now it’s a case of seeking out the right kind of supplier.

The bottom line

Gone are the days when how big an array to purchase and its expansion limits were so central to storage related investment decisions. Traditional technologies have evolved, and whole new architectural approaches have emerged in response to cloud computing demands in particular. As these increasingly find their way into fully supported commercial systems, it’s important to take time out and consider your options before moving forward with more of the same. We hope this short paper has given you a glimpse of what can be achieved with at least one of these options.

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we aim to help busy IT and business professionals get up to speed on the latest technology developments, and make better informed investment decisions.

For more information, and access to our library of free research, please visit www.freeformdynamics.com.

About Fujitsu

Fujitsu is the leading Japanese information and communication technology (ICT) company offering a full range of technology products, solutions and services. Approximately 162,000 Fujitsu people support customers in more than 100 countries. We use our experience and the power of ICT to shape the future of society with our customers.

Hyper-scale storage

ETERNUS CD10000 provides unlimited, modular scalability of storage capacity and performance at zero downtime for instant and cost efficient online access to extensive data volumes. Integrating open-source Ceph software into a storage system delivered with end-to-end maintenance from Fujitsu enables IT organizations to fully benefit from open standards without implementation and operational risks. Providing hyper-scalable object, block, and file storage up to more than 50 PetaBytes of data in a cost optimized way ETERNUS CD10000 is the ideal storage for OpenStack users, service providers for cloud, IT and telecommunication as well as media-broadcasting companies. Financial and public institutions with ever-growing document repositories, large scale business analytics / big data applications as well as organizations with comprehensive multimedia data can be served by ETERNUS CD10000 in an excellent manner.

For more information, please see http://www.fujitsu.com/eternus-cd

Terms of Use