FREQUENTLY ASKED QUESTIONS

These questions have been frequently asked on the ceph-users and ceph-devel mailing lists, the IRC channel, and on the Ceph.com blog.

IS CEPH PRODUCTION-QUALITY?

Ceph's object store (RADOS) is production ready. Large-scale storage systems (i.e., petabytes of data) use Ceph's RESTful Object Gateway (RGW), which provides APIs compatible with Amazon's S3 and OpenStack's Swift. Many deployments also use the Ceph Block Device (RBD), including deployments of OpenStack and CloudStack. Inktank provides commercial support for the Ceph object store, Object Gateway, block devices and CephFS with running a single metadata server.

The CephFS POSIX-compliant filesystem is functionally complete and has been evaluated by a large community of users. There are production systems using CephFS with a single metadata server. The Ceph community is actively testing clusters with multiple metadata servers for quality assurance. Once CephFS passes QA muster when running with multiple metadata servers, Inktank will provide commercial support for CephFS with multiple metadata servers, too.

WHAT KIND OF HARDWARE DOES CEPH REQUIRE?

Ceph runs on commodity hardware. A typical configuration involves a rack mountable server with a baseboard management controller, multiple processors, multiple drives, and multiple NICs. There are no requirements for proprietary hardware. For details, see Ceph Hardware Recommendations.

WHAT KIND OF OS DOES CEPH REQUIRE?

Ceph runs on Linux for both the client and server side.

Ceph runs on Debian/Ubuntu distributions, which you can install from APT packages.

Ceph also runs on Fedora and Enterprise Linux derivates (RHEL, CentOS) using RPM packages .

You can also download Ceph source tarballs and build Ceph for your distribution. See Installation for details.

HOW CAN I GIVE CEPH A TRY?

Follow our Quick Start guides. They will get you up an running quickly without requiring deeper knowledge of Ceph. Our Quick Start guides will also help you avoid a few issues related to limited deployments. If you choose to stray from the Quick Starts, there are a few things you need to know.

We recommend using at least two hosts, and a recent Linux kernel. In older kernels, Ceph can deadlock if you try to mount CephFS or RBD client services on the same host that runs your test Ceph cluster. This is not a Ceph-related issue. It's related to memory pressure and needing to relieve free memory. Recent kernels with up-to-date `glibc` and `syncfs(2)` reduce this issue considerably. However, a memory pool large enough to handle incoming requests is the only thing that guarantees against the deadlock occuring. When you run Ceph clients on a Ceph cluster machine, loopback NFS can experience a similar problem related to buffer cache management in the kernel. You can avoid these scenarios entirely by using a separate client host, which is more realistic for deployment scenarios anyway.

We recommend using at least two OSDs with at least two replicas of the data. OSDs report other OSDs to the monitor, and also interact with other OSDs when replicating data. If you have only one OSD, a second OSD cannot check its heartbeat. Also, if an OSD expects another OSD to tell it which placement groups it should have, the lack of another OSD prevents this from occurring. So a placement group can remain stuck "stale" forever. These are not likely production issues.

Finally, Quick Start guides are a way to get you up and running quickly. To build performant systems, you'll need a drive for each OSD, and you will likely benefit by writing the OSD journal to a separate drive from the OSD data.

HOW MANY OSDS CAN I RUN PER HOST?

Theoretically, a host can run as many OSDs as the hardware can support. Many vendors market storage hosts that have large numbers of drives (e.g., 36 drives) capable of supporting many OSDs. We don't recommend a huge number of OSDs per host

though. Ceph was designed to distribute the load across what we call “failure domains.” See CRUSH Maps for details.

At the petabyte scale, hardware failure is an expectation, not a freak occurrence. Failure domains include datacenters, rooms, rows, racks, and network switches. In a single host, power supplies, motherboards, NICs, and drives are all potential points of failure.

If you place a large percentage of your OSDs on a single host and that host fails, a large percentage of your OSDs will fail too. Having too large a percentage of a cluster’s OSDs on a single host can cause disruptive data migration and long recovery times during host failures. We encourage diversifying the risk across failure domains, and that includes making reasonable tradeoffs regarding the number of OSDs per host.

CAN I USE THE SAME DRIVE FOR MULTIPLE OSDS?

Yes. **Please don’t do this!** Except for initial evaluations of Ceph, we do not recommend running multiple OSDs on the same drive. In fact, we recommend **exactly** the opposite. Only run one OSD per drive. For better performance, run journals on a separate drive from the OSD drive, and consider using SSDs for journals. Run operating systems on a separate drive from any drive storing data for Ceph.

Storage drives are a performance bottleneck. Total throughput is an important consideration. Sequential reads and writes are important considerations too. When you run multiple OSDs per drive, you split up the total throughput between competing OSDs, which can slow performance considerably.

WHY DO YOU RECOMMEND ONE DRIVE PER OSD?

Ceph OSD performance is one of the most common requests for assistance, and running an OS, a journal and an OSD on the same disk is a frequently the impediment to high performance. Total throughput and simultaneous reads and writes are a major bottleneck. If you journal data, run an OS, or run multiple OSDs on the same drive, you will very likely see performance degrade significantly–especially under high loads.

Running multiple OSDs on a single drive is fine for evaluation purposes. We even encourage that in our 5-minute quick start. However, just because it works does NOT mean that it will provide acceptable performance in an operational cluster.

WHAT UNDERLYING FILESYSTEM DO YOU RECOMMEND?

Currently, we recommend using XFS as the underlying filesystem for OSD drives. We think `btrfs` will become the optimal filesystem. However, we still encounter enough issues that we do not recommend it for production systems yet. See Filesystem Recommendations for details.

HOW DOES CEPH ENSURE DATA INTEGRITY ACROSS REPLICAS?

Ceph periodically scrubs placement groups to ensure that they contain the same information. Low-level or deep scrubbing reads the object data in each replica of the placement group to ensure that the data is identical across replicas.

HOW MANY NICS PER HOST?

You can use one NIC per machine. We recommend a minimum of two NICs: one for a public (front-side) network and one for a cluster (back-side) network. When you write an object from the client to the primary OSD, that single write only accounts for the bandwidth consumed during one leg of the transaction. If you store multiple copies (usually 2-3 copies in a typical cluster), the primary OSD makes a write request to your secondary and tertiary OSDs. So your back-end network traffic can dwarf your front-end network traffic on writes very easily.

WHAT KIND OF NETWORK THROUGHPUT DO I NEED?

Network throughput requirements depend on your load. We recommend starting with a minimum of 1GB Ethernet. 10GB Ethernet is more expensive, but often comes with some additional advantages, including virtual LANs (VLANs). VLANs can dramatically reduce the cabling requirements when you run front-side, back-side and other special purpose networks.

The number of object copies (replicas) you create is an important factor, because replication becomes a larger network load than the initial write itself when making multiple copies (e.g., triplicate). Network traffic between Ceph and a cloud-based system such as OpenStack or CloudStack may also become a factor. Some deployments even run a separate NIC for

management APIs.

Finally load spikes are a factor too. Certain times of the day, week or month you may see load spikes. You must plan your network capacity to meet those load spikes in order for Ceph to perform well. This means that excess capacity may remain idle or unused during low load times.

CAN CEPH SUPPORT MULTIPLE DATA CENTERS?

Yes, but with safeguards to ensure data safety. When a client writes data to Ceph the primary OSD will not acknowledge the write to the client until the secondary OSDs have written the replicas synchronously. See How Ceph Scales for details.

The Ceph community is working to ensure that OSD/monitor heartbeats and peering processes operate effectively with the additional latency that may occur when deploying hardware in different geographic locations. See Monitor/OSD Interaction for details.

If your data centers have dedicated bandwidth and low latency, you can distribute your cluster across data centers easily. If you use a WAN over the Internet, you may need to configure Ceph to ensure effective peering, heartbeat acknowledgement and writes to ensure the cluster performs well with additional WAN latency.

The Ceph community is working on an asynchronous write capability via the Ceph Object Gateway (RGW) which will provide an eventually-consistent copy of data for disaster recovery purposes. This will work with data read and written via the Object Gateway only. Work is also starting on a similar capability for Ceph Block devices which are managed via the various cloudstacks.

HOW DOES CEPH AUTHENTICATE USERS?

Ceph provides an authentication framework called cephx that operates in a manner similar to Kerberos. The principal difference is that Ceph's authentication system is distributed too, so that it doesn't constitute a single point of failure. For details, see Ceph Authentication & Authorization.

DOES CEPH AUTHENTICATION PROVIDE MULTI-TENANCY?

Ceph provides authentication at the pool level, which may be sufficient for multi-tenancy in limited cases. Ceph plans on developing authentication namespaces within pools in future releases, so that Ceph is well-suited for multi-tenancy within pools.

CAN CEPH USE OTHER MULTI-TENANCY MODULES?

The Bobtail release of Ceph integrates the Object Gateway with OpenStack's Keystone. See Keystone Integration for details.

DOES CEPH ENFORCE QUOTAS?

Currently, Ceph doesn't provide enforced storage quotas. The Ceph community has discussed enforcing user quotas within CephFS.

DOES CEPH TRACK PER USER USAGE?

The CephFS filesystem provides user-based usage tracking on a subtree basis. RADOS Gateway also provides detailed per-user usage tracking. RBD and the underlying object store do not track per user statistics. The underlying object store provides storage capacity utilization statistics.

DOES CEPH PROVIDE BILLING?

Usage information is available via a RESTful API for the Ceph Object Gateway which can be integrated into billing systems. Usage data at the RADOS pool level is not currently possible but is on the roadmap.

CAN CEPH EXPORT A FILESYSTEM VIA NFS OR SAMBA/CIFS?

Ceph doesn't export CephFS via NFS or Samba. However, you can use a gateway to serve a CephFS filesystem to NFS or Samba clients.

CAN I ACCESS CEPH VIA A HYPERVISOR?

Currently, the QEMU hypervisor can interact with the Ceph block device. The KVM module and the *librbd* library allow you to use QEMU with Ceph. Most Ceph deployments use the *librbd* library. Cloud solutions like OpenStack and CloudStack interact libvirt and QEMU to as a means of integrating with Ceph.

Ceph integrates cloud solutions via `libvirt` and QEMU. The Ceph community is also looking to support the Xen hypervisor in a future release.

There is interest in support for VMWare, but there is no deep-level integration between VMWare and Ceph as yet.

CAN BLOCK, CEPHFS, AND GATEWAY CLIENTS SHARE DATA?

For the most part, no. You cannot write data to Ceph using RBD and access the same data via CephFS, for example. You cannot write data with RADOS gateway and read it with RBD. However, you can write data with the RADOS Gateway S3-compatible API and read the same data using the RADOS Gateway Swift-comptatible API.

RBD, CephFS and the RADOS Gateway each have their own namespace. The way they store data differs significantly enough that it isn't possible to use the clients interchangeably. However, you can use all three types of clients, and clients you develop yourself via `librados` simultaneously on the same cluster.

WHICH CEPH CLIENTS SUPPORT STRIPING?

Ceph clients–RBD, CephFS and RADOS Gateway–providing striping capability. For details on striping, see Striping.

WHAT PROGRAMMING LANGUAGES CAN INTERACT WITH THE OBJECT STORE?

Ceph's `librados` is written in the C programming language. There are interfaces for other languages, including:

- C++
- Java
- PHP
- Python
- Ruby

CAN I DEVELOP A CLIENT WITH ANOTHER LANGUAGE?

Ceph does not have many native bindings for `librados` at this time. If you'd like to fork Ceph and build a wrapper to the C or C++ versions of `librados`, please check out the Ceph repository. You can also use other languages that can use the `librados` native bindings (e.g., you can access the C/C++ bindings from within Perl).

DO CEPH CLIENTS RUN ON WINDOWS?

No. There are no immediate plans to support Windows clients at this time. However, you may be able to emulate a Linux environment on a Windows host. For example, Cygwin may make it feasible to use `librados` in an emulated environment.

HOW CAN I ADD A QUESTION TO THIS LIST?

If you'd like to add a question to this list (hopefully with an accompanying answer!), you can find it in the doc/ directory of our main git repository:

> https://github.com/ceph/ceph/blob/master/doc/faq.rst

We use Sphinx to manage our documentation, and this page is generated from reStructuredText source. See the section on Building Ceph Documentation for the build procedure.