Introduction To Guaranteed Quality Of Service

Quality of service (QoS) is a critical enabling technology for enterprises and service providers wanting to deliver consistent primary storage performance to business-critical applications in a multi-tenant or enterprise infrastructure. The type of applications that require primary storage services typically demand greater levels of performance than what is readily available from traditiownal storage infrastructures today. However, simply providing raw performance is often not the only objective in these use cases. For a broad range of business-critical applications, consistent and predictable performance are the more important metrics. Unfortunately, neither is easily achievable within traditional storage arrays.

> *"If your primary storage vendor does not have Storage QoS on its roadmap, now is the time to start demanding it."*
>
> Henry Baltazar
> Forrester Research

There is a large imbalance today between the performance and capacity resources within traditional storage systems. Capacity is plentiful and low cost; conversely, input/output per second (IOPS) are scarce and very expensive. From a provisioning perspective, performance and capacity are rigidly bound together, which only makes matters worse. This bind forces administrators to unnecessarily add storage capacity to increase the amount of IOPS available to a particular application. What results is a wasteful allocation of resources in an effort to overcome the limitations of existing storage architectures.

For service providers and enterprise IT, the promise of delivering storage resources predictably to a broad set of applications without worry has been nothing more than a pipe dream.

The History

QoS features exist in everything from network devices, to hypervisors, to storage. When multiple workloads share a limited resource, QoS helps provide control over how that resource is shared and prevents the noisiest neighbor (application) from disrupting the performance of all the other applications on the same system.

In networking, QoS is an important part of allowing real-time protocols such as VoIP to share links with other less latency-sensitive traffic. Hypervisors provide both hard and soft QoS by controlling access to many resources including CPU, memory, and network. QoS in storage is less common. If you seek out QoS within the storage ecosystem you will find that most approaches to storage QoS are "soft" – that is, based on simple prioritization of volumes rather than hard guarantees around performance.

Soft QoS features like rate limiting, prioritization, and tiering, are effective only as long as the scope of the problem remains small. When storage is deployed at scale these soft techniques quickly fail. In fact, these features are all "bolt-on" technologies that attempt to overcome limitations in storage architectures that were never designed to deliver QoS in the first place.

Guaranteed QoS: A critical component of the next generation data center

> *"Storage system capacity is no longer a top concern among the IT professionals I speak with. It's been replaced with 'How do I maintain top performance for a given application?' This is especially true in the virtual environment, where storage I/O is shared. With no guarantee of a specific performance level, mission-critical applications will not be virtualized."*
>
> George Crump
> Storage Switzerland

A quick look across today's storage landscape shows systems with a broad range of capacity and performance resources. On one end of the spectrum, disk-based systems have a high level of capacity and low level of performance. On the other end, flash architectures deliver a very high level of performance while requiring significantly less capacity (and at much higher cost). When viewed from the application perspective, the reality is that most application performance requirements fall somewhere in the middle of these two storage extremes.

In order to meet varying application performance requirements, the storage industry has responded by implementing caching or tiering schemes in front of traditional disk-based systems. These schemes apply complex algorithms and predictive methodologies that shuffle data to the right media at the right time to boost performance. Costly, complex, and reactive, this approach does little to bring you closer to the predictable performance required by mission-critical applications.

Solving for this disparity requires a more balanced pool of capacity and performance at the system level. From this starting point, a storage system can then deliver performance and capacity scaled independently to serve the unique needs of different applications. This ability to finely allocate capacity and performance resources separately from one another is a fundamental component of next generation data centers.

In these next generation infrastructures raw storage performance is important, but it is the predictable and consistent delivery of that performance which ensures every application has the resources required to run without variance or interruption. In servicing these workloads, IT must consider how well the underlying storage architecture will endure the following conditions:

- Unpredictable I/O patterns
- Noisy neighbor applications
- Constantly changing workload and application performance requirements
- Deduplication, compression, and thin provisioning processes
- Scaling of performance and capacity resources on demand

Where does "QoS" come from?

"Quality of Service," or "QoS", originated in the mid-1990s and referred to the overall performance quality experienced by end-users of a telecommunications network.

The term entered the storage realm about five years ago when SolidFire introduced a unique storage architecture specifically designed with the ability to control performance independent of capacity and deliver that performance predictably to thousands of applications within a single storage infrastructure. We now see QoS-like features popping up in the offerings of many storage vendors.

Server virtualization changed the way the world used computing, solving existing problems around inefficiencies, over-provisioning, and cost. But these advances were largely confined to compute and memory, and bypassed the seemingly unaddressable problems associated with storage: unr eliable performance and expensive resources. As a result, we only did half the job. Storage QoS helps with the other half.

Guaranteed QoS is not a feature — It’s an architecture

QoS is a system design choice that must be considered from the very beginning. True QoS delivers predictable performance natively, without having to optimize or organize data layouts to achieve it. Rate limiting, prioritization schemes, and tiering algorithms are all afterthoughts which attempt to overcome limitations in storage systems that were never designed to deliver predictable performance in the first place.

Being able to guarantee performance in all situations – including failure scenarios, system overload, variable workloads, and elastic demand – requires an architecture built from the ground up specifically to guarantee QoS. Trying to bolt QoS onto an architecture that was never designed to deliver performance guarantees is like strapping a jet engine to a VW Beetle. The wheels will come off just when you get up to speed.

The right storage architecture can overcome every predictability challenge by adhering to six core architectural requirements. Together, these six requirements enable true storage QoS and establish the benchmark for guaranteeing performance to every workload.

All-SSD architecture

- Enables the delivery consistent latency for every I/O

True scale-out architecture

- Linear, predictable performance gains as system scales

RAID-less data protection

- Predictable performance in any failure condition

Balanced load distribution

- Eliminate hot spots that create unpredictable I/O latency

Fine-grain QoS control

- Completely eliminate noisy neighbors, and guarantee volume performance

Performance virtualization

- Control performance independent of capacity and on demand

> “Quality of service should not be regarded as a feature that can simply be added to a storage product. QoS functionality that is bolted on after the fact tends to leave conditions in which performance is unpredictable and remains a non-starter for business-critical applications. Complete storage QoS requires [consideration and implementation] at the very core of storage product design.”
>
> Simon Robinson
> 451 Group

Hard QoS controls are defined by rigid terms such as IOPS and MB/s that are strictly enforced and produce predictable results regardless of system function or application activity.

Within the SolidFire platform, each volume is configured with minimum, maximum, and burst IOPS values that are strictly enforced within the system. The minimum IOPS provides a guarantee for performance, independent of what other applications on the system are doing. The maximum and burst values control the allocation of performance and deliver consistent performance to workloads. For the enterprise and service provider, SolidFire QoS enables SLAs around exact performance metrics and complete control over the customer’s experience. For infrastructure consumers, hard QoS delivers clear expectations around storage performance and the ability to deploy all tier 1 and tier 2 applications in the cloud with confidence.