Adaptive Quota Management

Global mobile traffic is forecast to grow from 99 exabytes per month in 2021 to 938 exabytes per month in 2026, a growth rate of 55% annually.* Video, online gaming and mixed reality applications are the main drivers behind that growth and will spur increased deployments of LTE-Advanced, LTE-Advanced Pro and 5G to meet the need. With many new applications, new devices and new use cases, traffic will be increasingly unpredictable in nature.

Accurately monetizing that surge in traffic and protecting the core network from excess control and signaling traffic while driving a consistent, predictable user experience is a significant challenge for mobile network operators. Meeting that challenge is key to maintaining customer loyalty, advocacy and a high Net Promoter Score (NPS).

ADAPTIVE QUOTA MANAGEMENT OVERVIEW

Enabled by the MATRIXX Technology Core, Adaptive Quota Management (AQM) is a unique, patented algorithm running on the MATRIXX Digital Commerce Platform (DCP). It is designed to provide highly accurate quota management, particularly when usage thresholds are being approached.

AQM has two primary purposes:

- The primary **monetization** purpose is to drive a superior end-user commercial experience by accurately and precisely managing balances and quota-granting at threshold points, eliminating erroneous reporting and denial of service.
- The primary **operational** purpose is reduction of the signaling load on the main monetization interface between the Packet Gateway (PGW) and Online Charging System (OCS) in the 4G Evolved Packet Core, and similarly between the Session Management Function (SMF) and Converged Charging System (CCS) in the 5G Standalone Core.

Not only will the monetization interface need to cope with increased quota management requirements driven by more traffic, but devices switching between 4G and 5G networks are expected to increase the signaling load by a factor of four.

*Source: Cisco

As quota (such as data megabytes) is requested and granted between the network node (PGW/SMF) and the MATRIXX DCP, AQM measures the rate at which a specific session is consuming a balance or asset. It then automatically adjusts the quota sizes and expiration times returned to the network to meet configured business and network requirements.

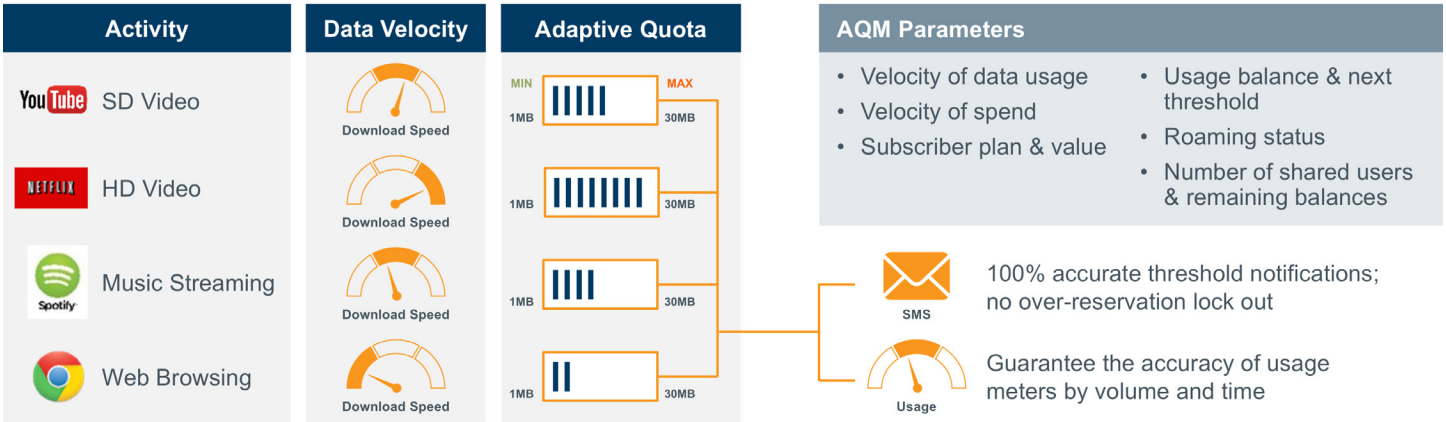
The MATRIXX-patented implementation of AQM enables operators to configure and automate timing accuracy curves for every type of session, specifying not only the required average message load for the session but also a geometric decay curve. This is followed as the notification threshold approaches, efficiently reducing quota sizes in a "fair allocation" scheme so that all sessions sharing the balance cooperate to hit the threshold without "quota starvation" or hoarding. It is the only quota management solution that automatically adapts the quota size decisions based on the actual consumption rates of all sessions involved.

ADAPTIVE QUOTA MANAGEMENT BENEFITS SUMMARY

AQM is uniquely able to provide accurate, timely usage threshold notifications, even across shared enterprise balances. It achieves this without resorting to tiny quotas that will escalate network signaling loads out of control or large quotas that will result in erroneous reporting and denial of service. Real-world scenarios have shown an almost 70% reduction in Transaction per Second (TPS) loads at the OCS/PGW Gy interface. It ensures that no application in a single balance environment or device in a shared balance environment either hogs or starves other applications or devices of available quota. Analogous to cruise control in cars, the automated, adaptable nature of AQM provides a highly accurate, predictable and equitable means of quota consumption at threshold points. The dual benefits of maintaining that consistently accurate end-user experience while reducing network loading amidst highly unpredictable usage patterns make AQM an operational necessity.

Figure 1: An Overview of Adaptive Quota Management

- AQM learns and adapts quota sizes and validity times for each session
- Reduces load on core network and OCS/CCS
- Provides a better customer experience than using fixed quota amounts



ADAPTIVE QUOTA MANAGEMENT DETAILED OVERVIEW

AQM is a standard feature of the MATRIXX Digital Commerce Platform and is configurable through the MyMATRIXX Graphical User Interface.

The challenge of managing dynamic quota allocation is outlined in the tables below. While some approaches advocate increasing and hard configuring a static, large quota block size to minimize the signaling load across the Gy interface between the OCS and PGW, diminishing returns apply above a certain block size.

AQM dynamically and automatically adjusts quota size, reservation size and validity times by actively monitoring balance and quota consumption velocity and thresholds. It optimizes network communication and the accuracy of balance queries and threshold notifications, and is self-adjusting for different services, minimizing the need for service-specific configuration.

Figure 2: The Challenge of Managing Dynamic Quota

Session Size	% of Sessions in Sample	% of Total Volume per Sample (in MB)
< 1MB	61%	0.5%
1MB – 5MB	15%	2.5%
5MB – 10MB	7%	3%
10MB – 50MB	12%	22%
50MB – 100MB	3%	18%
> 100MB	2%	54%

- Session samples were taken over 24-hour period across both 3G and 4G network
- Pattern shows typical trend of large percentages of small connections using little data
- Small percentage of connections using large amounts of data
- Total volume of data in sample was over 150TB in under 24 hours

Quota Block Size	Total Gy Interactions for Sample	Reduction in Gy by Increasing Quota
1MB	100%	N/A
2MB	71.67%	28.33%
5MB	54.12%	45.88%
10MB	48.73%	51.27%
50MB	44.55%	55.45%
100MB	44.13%	55.87%

Increasing quota size from 1MB to 2MB and up to 10MB has an immediate return in terms of reducing Gy load between network and OCS. However, larger increases in quota have less advantages.

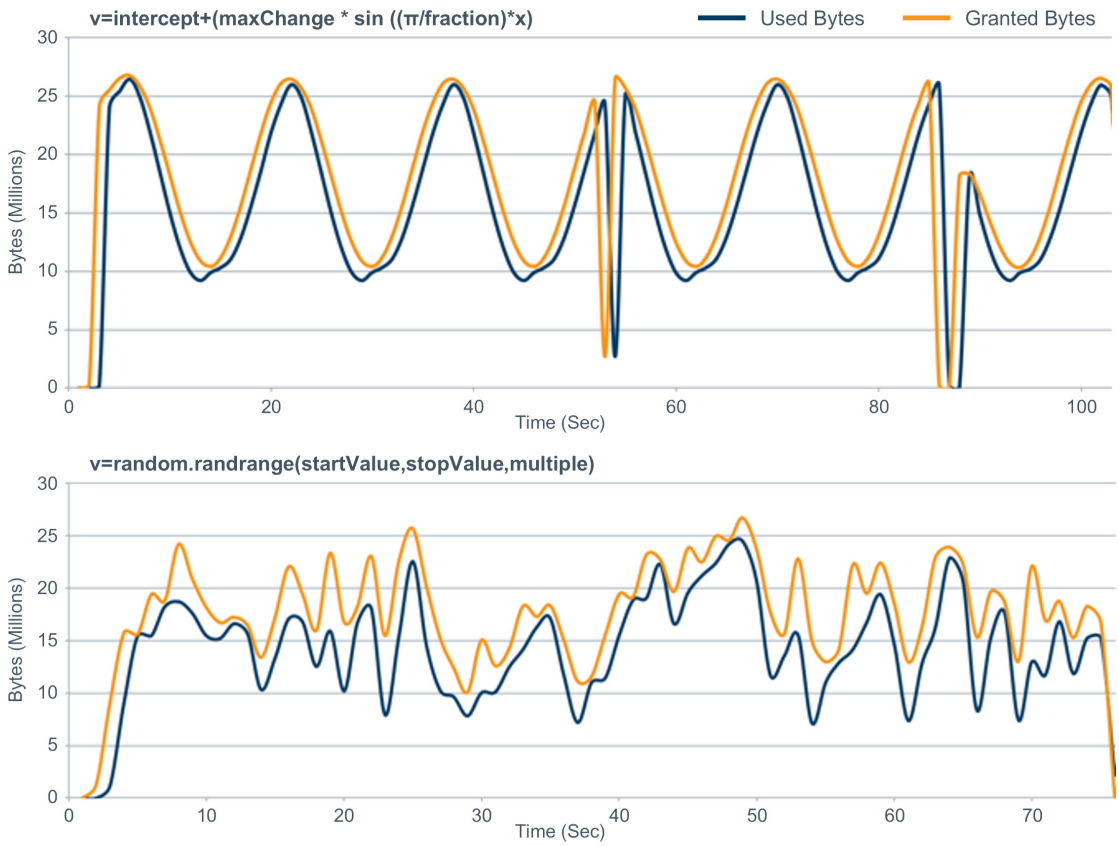
In this sample, doubling of quota size from 50MB to 100MB has almost no benefit.

A series of requests for a single service context in an AQM session is called a flow. There can be multiple flows per context and per session. For each flow, the MATRIXX Charging Application dynamically adjusts the size and validity times of reservations and quotas.

AQM allocates the maximum quota to each flow considering the current quota consumption velocity by the flow against balances, the global quota consumption from all users and devices against balances and the time that the asset needs to be available for use. In addition to quota consumption velocity, AQM considers pricing when allocating quota.

In test scenarios, when using different patterns of consumption (sinusoidal and random), the accuracy of AQM in matching the quota used to the quota granted has been clearly proven. This exactness is maintained per-flow right up to key threshold points.

Figure 3: Test Scenarios Showing the Impact of AQM on Sinusoidal and Random Traffic Consumption Patterns



A balance may be unshared or shared.

- **Unshared balance** – An unshared balance is reserved by exactly one flow and AQM grants quota so that a threshold is reached or crossed to within one beat of the actual threshold amount.
- **Shared balance** – A shared balance is reserved by multiple flows; the behavior of one flow can affect the threshold accuracy of other flows. In this case, AQM grants minimal quotas near each threshold so that, collectively, the flows cross the threshold at approximately the same time.

During AQM, if a balance is reserved by two or more flows, it is a shared balance. If all flows reauthorize in one message, the first flow appears to have a reserved amount of zero and the balance is considered unshared.

The MATRIXX Charging Application calculates the individual session per-balance velocity and the global per-balance velocity. These values impact the AQM algorithms and are recalculated any time session activity reports usage or significantly affects the session. MATRIXX Charging Application measures balance velocity by sampling the gross amount associated with each reservation and then computing the change in that value over time.

AQM monitors quota velocity (average used quantity over time) and balance velocity (the change in the gross balance over time). It also tracks previous velocities and uses that information to determine a weighted moving average to lessen the impact of usage volatility. Note that velocity is not controlled by AQM, but rather by factors such as the following:

- Type of content (such as video, audio or text)
- Radio Access Technology (such as 4G or 5G)
- Connection strength
- QoS (such as throttled or unthrottled)

AQM CONFIGURATION

The allocated quota is determined by a combination of service-type definition attributes, balance template attributes and metrics of historical quota and balance consumption.

Service configuration attributes control the quantity (in bytes) and validity (in seconds) of the quota to be authorized. Balance template attributes control the quantity of the units configured in the template (such as Kbytes or dollars) and validity (in seconds) of reservations against the balance. Service attributes take into consideration only the quota velocity (for example, HD video usually gets larger quota than email) while the balance template attributes take into consideration both quota velocity and price. For example, a cheaper offer usually gets a larger quota than a more expensive offer. If a maximum reservation is $1.00, 20MB at $0.05/MB could be allocated, but only 10MB at $0.10/MB.

AQM USE CASES

The MATRIXX DCP flexibly supports the monetization of consumers, enterprises, SMBs and IoT devices on the same platform. As a key technology core-enabled feature, AQM can be deployed against any type of device or user.

Single Balance Consumer

A typical smartphone consumer will use quota at wildly differing rates per application. As an example, in Figure 1, HD Video, SD video, music streaming and web browsing will all behave differently, though this behavior is invisible to the user.

As previously outlined, artificially and statically configuring small quota blocks in an attempt to hit threshold points consistently per application places an unnecessary and unacceptable signaling load on the network interface. Similarly, setting quota block sizes too high results in "hungry" applications, such as HD video consuming quota more rapidly than others, hitting a threshold point first and being denied any further quota. If this occurs while other applications continue to work normally, it results in a poor customer experience. Managing all of the application flows on a particular device in a consistent manner, such that they are all automatically and independently stepped down in quota usage and reach the threshold point at the same time, massively reduces inaccuracies, denial of service and "bill shock."

Shared Balance SMB/Enterprise

Enterprise/SMBs often have complex organizational hierarchies with many parent-child or departmental peer relationships evident. Maintaining overall balance and quota usage, consistency, allocations and tracking is a real challenge for any real-time commerce engine. The MATRIXX DCP, enabled by its technology core, can support highly scalable, hierarchical, shared enterprise balances supporting many tens of thousands of devices. Alongside that foundational scale, AQM has a key role to play in ensuring that quotas allocated across that hierarchy are gracefully and consistently managed and adapted at threshold points so that no device, department or group of devices are unfairly starved of quota if there is quota available in the overall "pool."

ABOUT MATRIXX DIGITAL COMMERCE

MATRIXX Digital Commerce is the industry's leading cloud native rating and convergent charging solution. Architected with the performance and resiliency of a network function and the configurability of an IT application, MATRIXX DCP unifies IT and Networks to provide operators the agility to operate at web scale. With its API-first design, lightweight, no-code configuration, and microservices architecture, MATRIXX Digital Commerce is easy to configure, fast to deploy and capable of serving multi-network environments from a single, extensible platform. Massively scalable and highly efficient to operate, MATRIXX DCP enables operators to successfully automate operations, monetize new services and hyper-scale offerings, all at web-speed.

matrixx.com