Understanding Throttling

https://campus.barracuda.com/doc/100370668/

Many vendors, including Barracuda Cloud-to-Cloud Backup, use Microsoft Graph to access data from a variety of Microsoft cloud services. While Microsoft Graph is designed to handle a large number of requests, throttling can occur to ensure the service remains functional. Throttling is a mechanism that Microsoft Graph API uses to limit the number of requests or the rate at which requests can be made to the API. This is implemented for several reasons:

- Protect First-Party Experience – Throttling ensures Microsoft customers receive a first-class experience while using Microsoft 365.
- Fair Resource Allocation – Throttling ensures that all users and applications get fair access to Microsoft Graph resources. Without throttling, a single application could monopolize the API, causing a poor experience for other users and applications.
- Load Balancing – Throttling helps distribute the load on Microsoft's servers evenly. It prevents overloading of their systems, which could lead to service disruptions or slowdowns.
- Protection from Abuse – Throttling is a crucial security mechanism to protect against abuse or malicious activity. It helps prevent denial-of-service (DoS) attacks and other forms of abuse that could compromise the API's reliability.
- Performance and Efficiency – Throttling helps optimize API performance by ensuring that resources are used efficiently. It encourages developers to write more efficient code and reduces unnecessary load on servers.

There are multiple types of throttling that Microsoft utilizes at the Azure region level, the application level, the tenant level, and the user level (eg, a given Mailbox/OneDrive/Team/Site). The two primary types of throttling typically seen are:

- Application Throttling – This limits the number of requests that can be made by an application within a specific time frame. It is typically set at a reasonably high rate to accommodate most legitimate usage.
- User Throttling – This limits the number of requests a specific user or user identity can make within a specific time frame. This helps prevent a single user or application from monopolizing resources.

The specific limits for application and user throttling can vary based on multiple factors and are different for each data source (Exchange, OneDrive, SharePoint and Teams). These limits are subject to change by Microsoft, and they do change regularly. When this happens, it is necessary to update Cloud-to-Cloud Backup accordingly. Barracuda Network employs all strategies at our disposal to ensure your backup jobs are done quickly and efficiently within the guidelines enforced by Microsoft. Barracuda Networks is constantly evolving to improve application performance.

For more information on why Microsoft throttles Graph API traffic, see the Microsoft article https://learn.microsoft.com/en-us/graph/throttling.

For more information on how to change the Microsoft Exchange Web Service (EWS) throttle policy

limits for the Exchange data source, see How to Configure EWS Throttle Policy Limits.