# GigaFlow Sizing Document

This documents aim is to provide the engineer with sufficient information that they can right size an installation.

Although GigaFlow will run on almost any hardware that runs Java and Postgres, we would recommend a specification above the minimum requirements. This provides not only capacity to grow the installation but also ensures that the end user gets the most out of the software’s capabilities (i.e. faster reporting).

When sizing a server, the following items should be considered:

**Ingest Rate**: Cumulative number of flows per second (flows/s) being sent to the server.

**Device Count**: Number of devices sending flows to the server.

**Hardware Available**: Cumulative performance of CPU, RAM, Disk IO and Disk Space.

**Storage Duration**: How long the user wants to keep data for.

**Reporting Performance**: What queries the user wants to run and what is an acceptable response time for requests.

**Features/Functionality**: What features will be enabled and to what scale.

## Ingest Rate

The number of flows (Netflow/SFlow/JFlow/IPFIX/Syslog) being parsed by GigaFlow and is the largest scaling factor to be considered during deployments. The more flows/s that need to be ingested then the more work the server has to do. The more flows/s stored to disk also impacts reporting time as more data has to be accessed.

The flows/s rate can be seen in the GigaFlow UI:“System->System Status->Average Flows/s”

As well as the raw flow rate, the make-up of the flows can also have an effect. i.e. if the flows are from an internet facing device, this may generate a large number of “Device/Source IP/Destination IP” tuples for the software to track, even on small networks this can mandate an increase in the amount of RAM assigned to the GigaFlow process.

The number of tracked IP Addresses can be seen in the GigaFlow UI:“System->System Status->IPs” which shows how many IPs are currently being tracked in RAM and GigaFlow UI:“System->System Status->IP Window” which shows window size of the IPs cache. i.e. how long Ips are held in RAM for.

## Device Count

The device count is important as internally GigaFlow scales out the processing of flows on a per device basis. Devices are assigned to a CPU thread and the server must be suitably spec’d and configured so that all handling of that device flows at ingest time (and any other devices handled by the same core) shouldn’t overload any core.

The number of threads is set in “System->Receivers->Netflow Process Threads” button on top right of the page. This should never be more than ½ the number of available cores on the server.

The table at the bottom of the “System->Receivers” page shows the number devices, type of data and stats on each CPU thread.

The device count is also important at report time as all tables are keyed by device. When running reports, including a device filter can give significant performance improvements.

The following chart shows the relationship between device count and supported number of flows (total) recommended on a single GigaFlow instance.

## Hardware Available

The minimum hardware available to GigaFlow for the aforementioned device/flow rates should be at least equivalent to our Viavi GigaFlow appliance:

 CPU: 2 X Intel Xeon Silver 4214 2.2GHz (12 core per CPU)

RAM: 12x8GB=96GB

OS Drive: 2 X 960GB SSD in RAID1 (804GB formatted)

RAID Controller: 9361-4i Single

DATA Drive RAID: 10x8TB SAS HDD =  80T (in RAID6), 58.2TB after formatting

By Default, the server should use 8 cores for Postgres queries, 5 cores for flow receivers and the remaining for the OS, main Java process and other processing requirements. RAM is split with ¼ for Postgres, ¼ for GigaFlow and ½ for OS file caching (Postgres effective cache size settings).

## Storage Duration

Storage duration (how long to keep live data for on the disks) has 2 major impacts

1. The amount of disk space required should increase linearly with the number of days to keep data for (if the flow rate is stable).
2. The number of tables to be maintained. This has an impact on house keeping duties such as Rollups but also impact at report creation as more tables have to be queried or pruned from a query.

By default, GigaFlow keeps data based on device specific rollup schedules. For high flow devices tables will be created every hour, for smaller devices they may be kept in up-to 4-hour tables.

When data is older than the “Forensics Rollup Age” period (4 days by default), the low flow device tables will be grouped into single 24-hour duration tables to help maintain a manageable total table count.

It is recommended that the table count be kept below 15,000 tables so that query planning does not become excessive at 60,000 reporting will fail due to too many tables and the retention period should be reduced. The expected table count based on the current device/flow rates can be found at the GigaFlow UI:“System->Global->Storage-> Max Forensics Storage” setting value.

## Reporting Performance

This is impacted by the following factors:

1. Filters applied
2. Volume of data to process
3. Number of returned results
4. Hardware spec to deliver the report

All reporting in GigaFlow is handled by the Postgres database. It is for this reason that we give over most of the hardware resources to it so that reporting time can be as short as possible, delivering a good user experience, however, there will always be a trade-off between query performance, data volume and hardware costs.

## Features/Functionality

GigaFlow has a large number of configurable features and functions which may impact the overall performance of the system or mandate the increasing of available hardware resources.

As such, there is no way to exactly say what the performance of any given server will be (as the make-up of the flows will be different between servers) but we do know which features/functions require which hardware resource. The following table shows the top line functionality in GigaFlow, what hardware resource they are affected by, which factor to be aware of and any recommended (potentially unenforced) limits.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Affects** |  |  |
| **Function** | **CPU** | **Memory** | **Disk** | **Variable Load Factors Outside of Device and Flow Count** | **Recommended Limit (May require more RAM)** |
| Application Identification | Medium | Low | None | Number of additional applications defined (Complex or otherwise) | 50 |
| Deduplication | Medium | Low | None | Nothing configurable | N/A |
| IP Discovery | High | High | None | Number of unique IP addresses \* how many devices they have been seen on | 1,000,000 per hour |
| Traffic Groups | Low | Low | None | Number of Traffic Groups | 2000 |
| First Packet Response | Medium | Low | None | Number of monitored subnets and services monitored | 10 subnets, 100 services per subnet |
| Black listing | High | Low | None | Number of blacklists to monitor for | 30000 |
| Server Discovery | High | High | None | Devices and servers to monitor | N/A |
| Summary Builder | High | Low | Low | Number of device interfaces | 5000 |
| PTC Flow Creation | High | High | None | Number of (client, server, appId, device) tuples | 1/4 normal flow capacity |
| Syslog Flow Creation | High | High | None | Number of (client, server, appId, device) tuples | 1/4 normal flow capacity |
| Apex Flow Creation | High | High | High | Number of (client, server, appId, device) tuples | 1/4 normal flow capacity |
| Performance Overviews | High | Low | Low | Nothing configurable | N/A |
| Performance Traffic Groups | High | Low | Low | Number of Traffic Groups | 2000 |
| Syn Monitoring | High | Low | None | Nothing configurable | N/A |
| Profiling | High | Low | None | Number and complexity of profiles | 100 |
| Flow Storage | Low | Medium | High | Nothing configurable | See Chart |
| User Mapping | High | Low | None | Number of monitored users | 10,000 |
| Syslog Parsing | High | None | None | Parser complexity and string length | Parser will inform user during configuration  |