SmartHub.ai

SmartHub INFER IoT Center

v3.0.0

Server Deployment Sizing Guide

You can find the most up-to-date technical documentation at: https://www.smarthub.ai/

SmartHub Inc.
4332 Holt Street, Union City, CA, 94587, USA
www.smarthub.ai

Table of contents

1 Introduction

The SmartHub INFER IoT Center Sizing Guide captures the sizing guidelines for SmartHub INFER IoT Center for the number of devices managed and the use of components such as device management, metrics ingestion, alert generation, notifications, device updates, and command execution.

This guide also provides some best practices and server-side tuning parameters to fine-tune the environment for a specific data set and feature usage on a large scale.

SmartHub has tested INFER IoT Center with 15000 devices (2500 Gateways + 12500 Things). The details are listed in the following table.

The SmartHub INFER IoT Center offers an excellent control plane solution over Kubernetes to manage, monitor, update, and troubleshoot an enterprise IoT infrastructure and drive its operational efficiency. It supports a varied range of gateways and sensors. At the time of the writing installing SmartHub INFER as a Kubernetes cluster is not offered as a self service.

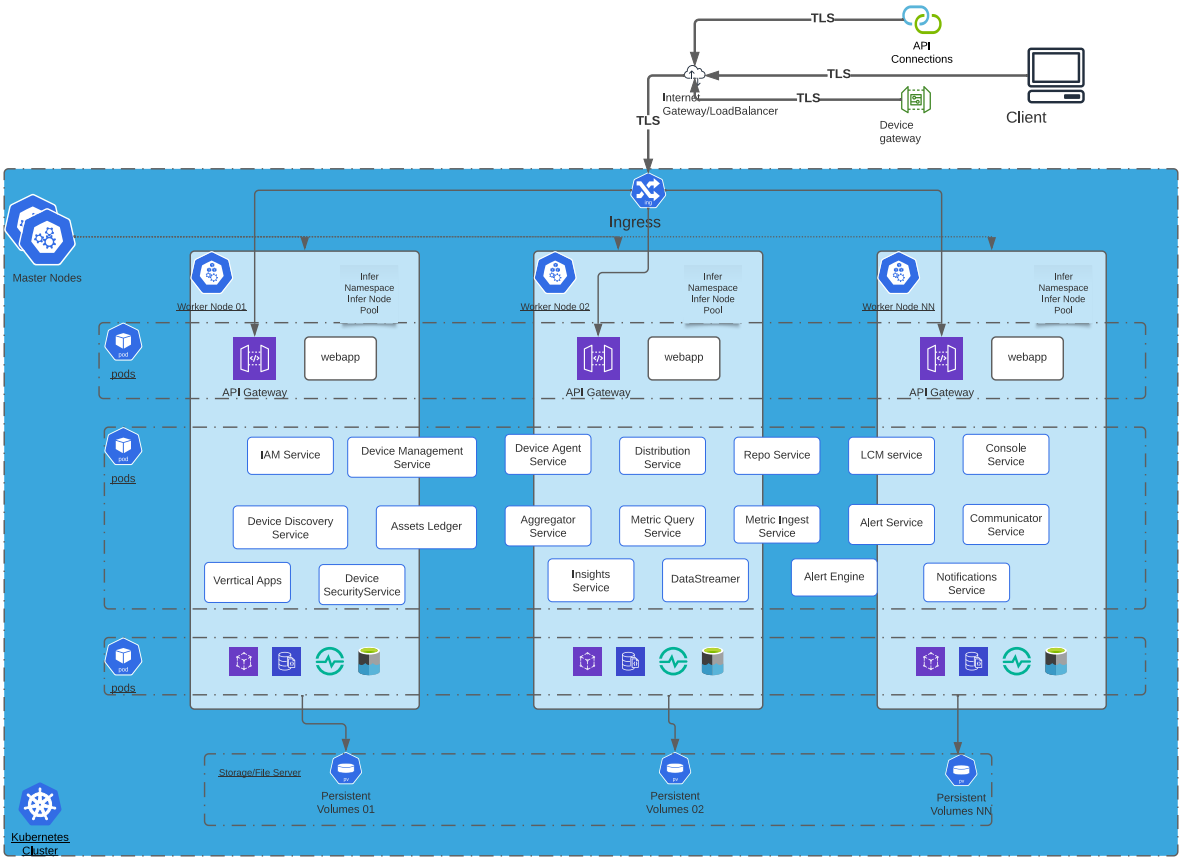

Figure 1-1. SmartHub INFER IoT Center 3.0.0 Deployment Diagram

2 Key Terminology

The following key terminologies are used throughout this guide.

2.1 Gateway versus Thing

Apart from the gateways that send the data to the SmartHub INFER IoT Center server, there are sensors that read the data from the physical world and send them to the gateways from time to time. These sensors are called Things. The SmartHub INFER IoTC Agent software residing at the gateway device aggregates the data and sends it to the SmartHub INFER IoT Center server.

2.2 What Are Metrics

Metrics are the data sent by the gateway devices to the SmartHub INFER IoT Center server. They are the parameters that are under observation at a given time. The values are ingested in the database and are displayed in a graphical representation in the SmartHub INFER IoT Center console.

2.3 What Is a Metrics Ingest Interval

Metrics Ingest Interval is the interval at which the gateway devices send metrics to the server. The default interval value is 5 minutes. However, the recommendation for Metrics Ingest Interval can vary depending on the scale of devices added to the system and the number of metrics sent to the server.

2.4 What Is a Sampling Frequency

There are several sensors connected to a gateway. These sensors send data to the gateway every 15 or 20 minutes. This interval is known as Sampling Frequency. However, the aggregation of this data is done by the gateway and sent at every Metrics Ingestion Interval set at the SmartHub INFER IoT Center server.

3 Basis of Sizing

The sizing guideline for SmartHub INFER IoT Center has been arrived at after performing the following tests.

1. Ran multiple small experiments for specific feature sets that were provided by customers as feedback. These tests are called workload models and are referred throughout this document.
2. Set up a large-scale test environment with 15000 enrolled virtual devices including gateways and things. Performed regular user operations and monitored the performance over an extended time.

The following workload model table lists the number of devices enrolled, the number of metrics or devices ingested, the number of alert definitions, the number of campaigns, and the number of commands run.

Test Unit	Value	Comments
Gateways	2500	
Things	12500	
Metric Point	10	30 data points per device per interval
Sampling Frequency	3	
Ingestion Interval	1 Hour	
Alert Definitions	20	
Alerts per Day	750	32 alerts per hour
Notifications	1500	
Campaigns	1	10 MB package
Commands	100 file uploads per day	2 MB file size

The following features were covered in the tests that were simulated at scale:

- SmartHub INFER IoTC Agents ingesting metrics to the SmartHub INFER IoT Center Server.
- Monitoring alerts generated in the system.
- Monitoring notifications generated in the system.
- Running campaigns using real agents.
- Running commands to fetch logs from devices.

4 Device Template Recommendations

Device templates are created to group different types of sensors and gateways that are used to report metrics to the SmartHub INFER IoT Center server.

There are certain limits set to these templates to help you design and group the devices in an effective way.

Each organization can have up to 100 templates. There is no limit for the number of devices per template. There can be a maximum of 1000 devices in an organization and a maximum of 100 templates in an organization.

5 Sizing Work Models

The sizing work model is determined by a baseline of feature usage and the number of devices used in the system. If the baseline of the feature usage varies, then it increases the resource usage of the underlying prescribed infrastructure.

SmartHub INFER deployment is purely a Kubernetes deployment. Each of the resources listed in the following table is divided equally among the kubernetes worker Nodes(VM / bare metal) to create a high availability setup. The sizing can be considered the bare minimum

Devices (gateways and things)	CPU Cores	RAM(GB)	Storage(GB)	Nodes
<= 1000	10	40	768	5
1001–5000	12	48	1024	6
5001–10000	20	80	2048	7
10001–20000	32	128	4096	8
> 20000	Contact Support	Contact Support	Contact Support	Contact Support

We recommend to keep some additional compute resources reserved to enable Kubernetes pod horizontal auto scaling.

To resize the environment for the number of devices in the sizing work model, contact SmartHub Support.

5.1 Assumptions

1. Size doesn’t include video or image processing/analytics
2. The sizing is not shareable and is exclusive to SmartHub INFER deployment only
3. Edge sizing is not considered and will depend on the use cases.
4. Doesn’t take into account any sizing needs for Kubernetes control plane nodes, app monitoring applications, log monitoring, additional nodes for upgrades, backups, DR etc.

6 Peak Load Recommendations

For handling peak load, we recommend enabling node pool scaling and pod auto scaling in kubernetes. Its recommended to have another 3 or more Kubernetes worker nodes to handle scale at peak load times as a part of the Kubernetes node auto scaling for the pod to auto scale.

7 Metrics Retention

SmartHub INFER Data Lake provides cross domain reporting between various data points and stores Device Metrics as Time-series data. This section explains the retention periods for the time series data. The Data Lake supports storing Metrics for time frames that can span beyond a year. The older the data becomes, the lesser the resolution becomes. In other words, time series data is stored at the maximum resolution for the most recent data. As the data ages, the number of data points per time range reduces by replacing high resolution data with mean values at a lower resolution.

Here is a table of reports for various time spans:

Data lake time ranges	Resolution (per hour/metric)	Resolution (per day/metric)
<= 30 days	Full	Full
31 – 90 days	10	240
91 – 180 days	5	120
181 – 365 days	0.5	12
> 365 days		1