# GCUT Dataset

## Google Cluster Usage 2011 traces (version 2.1) : ~ 41GB compressed

Data path in the cluster: /rdata/datasets/gcut\_dataset

* Github: [cluster-data/ClusterData2011\_2.md at master · google/cluster-data](https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md)
* Google AI Blog: [More Google Cluster Data](https://ai.googleblog.com/2011/11/more-google-cluster-data.html)
* **Google cluster-usage traces: format + schema** (Revised 2014-11-17 for trace version 2.1) : [Google cluster-usage traces format schema 2014-11-17 external.pdf](https://drive.google.com/file/d/0B5g07T_gRDg9Z0lsSTEtTWtpOW8/view)
* Towards understanding heterogeneous clouds at scale: Google trace analysis: <https://www.pdl.cmu.edu/PDL-FTP/CloudComputing/ISTC-CC-TR-12-101.pdf>
* GloudSim: Google trace based cloud simulator with virtual machines (Paper 2014): <https://www.mcs.anl.gov/papers/P5017-0913.pdf>

## Google Cluster Workload Traces 2019 (Version 3): ~ 2.4TiB compressed

The workloads running on **eight** Google Borg compute clusters for the month of May 2019.

* Google Cluster Workload Traces 2019: <https://research.google/tools/datasets/google-cluster-workload-traces-2019/>
* **Google cluster-usage traces v3:** [Google cluster-usage traces v3.pdf](https://drive.google.com/file/d/10r6cnJ5cJ89fPWCgj7j4LtLBqYN9RiI9/view)
* ClusterData 2019 traces (Github): <https://github.com/google/cluster-data/blob/master/ClusterData2019.md>
* Large-scale cluster management at Google with Borg (Paper): <https://research.google/pubs/pub49065/>
* Borg: the Next Generation (Paper): <https://research.google/pubs/pub49065/>
* BigQuery: <https://cloud.google.com/bigquery/>
* Google AI Blog: [Yet More Google Compute Cluster Trace Data](https://ai.googleblog.com/2020/04/yet-more-google-compute-cluster-trace.html)

*Note: If you are able to use* ***big-query****, try* ***Google ClusterData 2019****, and if you* ***need to work offline*** *with the dataset,* ***try the 2011 version.***

**I suggest for you to try the real data provided by Google. The Google workload traces are collected from large cloud systems (over than 12,500 compute nodes) during 29 days. The traces consist of different types of over than 25,000,000 tasks belong to about 930 users. Real workload traces can provide a very high level of realism when used directly in performance evaluation experiments.** [**link**](https://www.researchgate.net/post/Where-can-I-find-real-workload-traces-of-VMs-of-Cloud)