# DARTS: The Data Analysis Remote Treatment Service

31 May 2023


## Summary

We present the Data Analysis Remote Treatment Service (DARTS), an open-source remote desktop service that launches on-demand virtual machines in the cloud, and displays them in a browser. The released environments can be used for e.g. scientific data treatment. DARTS can be deployed and configured within minutes on a server, and can run any virtual machine. The service is fully configurable, supports GPU allocation, is scalable and resilient within a farm of servers. DARTS is designed around simplicity and efficiency. It targets laboratories and facilities that wish to quickly deploy remote data analysis solutions without investing in complex hypervisor infrastructures. DARTS is operated at Synchrotron SOLEIL, France, in order to provide a ready-to-use data treatment service for X-ray experiments.


## Statement of need

Synchrotron radiation facilities and other large-scale research facilities generate increasingly massive and complex amounts of data due to the nature of their experiments. This trend, referred to as the "data deluge," (Wang, Steiner, and Sepe 2018) is closely linked to the evolution of technological bricks such as detectors, storage, network, and computing capability.

To overcome this challenge, a sensible solution is to provide suitable software on powerful computers with an interactive remote access without the need for data transportation. By doing so, researchers can efficiently access and analyse their data without requiring expensive local hardware or software. Data analysis is a vital preliminary step in the production of scientific publications, which are the actual metric upon which research facilities are evaluated in their societal impact.

While the Jupyter ecosystem (Kluyver et al. 2016; Randles et al. 2017) is now widely used for scientific data analysis, it still requires users to have basic

knowledge of commands and scripting, and does not allow to launch full GUI applications. Alternatively, a number of commercial solutions exist, such as Amazon WorkSpaces (Amazon Web Services 2023), FastX (StarNet Com. Santa Clara CA 2023) and NX/NoMachine (NoMachine 2023). Other community related software exist, such as the VISA platform (Caunt Stuart 2021), the ISIS Data Analysis as a Service (Frazer Barnsley 2016), and the CoESRA service (Guru et al. 2016), but none of them is fully open-source, easily installable and deployable.

The Data Analysis Remote Treatment Service (DARTS) is a lightweight, on-demand, cloud service to instantiate and display ready-to-use complete scientific software environments.

# Implementation

The conceptual design of the Data Analysis Remote Treatment Service (DARTS) relies into the following sequential steps:

1. Identify a user and computing requirements from a web form (landing page).
2. Launch a copy of a master virtual machine.
3. Display it in a browser.

The DARTS service starts from the landing page, in which a user feeds information (credentials, computing requirements), and selects one of the available environments (from the `machines.conf` file). This information is collected by the main script (`qemu-web-desktop.pl`) which imports the main configuration `config.pl` and takes care of the whole service steps (instantiation, monitoring, self-cleaning). A snapshot of the selected master virtual machine environment is created, to hold user-level changes in the instance. It is then started and attached to the QEMU embedded VNC server. Start-up configuration script can be injected via `virt-customize` (Jones 2011) to be executed during the boot process. A websocket is exposing the internal VNC port as a URL, and displayed with noVNC. The result page is generated with the proper URL for the user to connect. The performance of the virtualization layer reaches native speed for both CPU and GPU, as well as disk and network.

Relying on a steady software stack (Apache2, Perl, QEMU) with limited dependencies, DARTS is easy to deploy and operate. In practice, the only DARTS-related maintenance action consists in adding or updating the virtual machines. The simplicity of DARTS only requires a fraction of a single staff for its administration.

## Research applications

DARTS is especially suited for small to medium research laboratories and facilities willing to quickly deploy a remote data analysis infrastructure, with minimal maintenance.

At the Synchrotron SOLEIL, the service has been operated continuously since 2020 for our users on two servers equiped with GPU's (Farhi 2023). Our current environments in production are a default Debian system holding X-ray data treatment software (currently 631 scientific applications and libraries), a reduced system meant to be distributed to the users as they leave the facility, and a Windows 10 system with commercial software. Our Debian images are built automatically via a set of shell scripts (Picca and Farhi 2022). This choice is meant to minimize our maintenance. These images mount a persistent user folder (also accessible via a JupyterHub service), as well as the experimental data storage via NFS, CIFS/Samba and SSHFS. In addition, information from the authentication service (LDAP) is used to customize each instance and install specific files and applications on top of existing master virtual machines.

## Author contribution statement

Conceptualization, coding, development and paper writing by Emmanuel Farhi.

## Acknowledgements

We thank the members of the Data Reduction and Analysis Group at Synchrotron SOLEIL, and more particularly Frédéric-Emmanuel Picca for his continuous support during the development of this project. We also thank Roland Mas, from the GNURANDAL company, for the Debian packaging. This project has received support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957189 "BIG-MAP" (Tejs Vegge 2020).

## References

Amazon Web Services, Inc. 2023. "Amazon Workspaces." https://aws.amazon.com/fr/workspaces.

Caunt Stuart, et al. 2021. "Virtual Infrastructure for Scientific Analysis." https://visa.readthedocs.io.

Farhi, Emmanuel. 2023. "DARTS: The Data Analysis Remote Treatment Service." https://data-analysis.synchrotron-soleil.fr/qemu-web-desktop/.

Frazer Barnsley, Tom Griffin, Brian Matthews. 2016. "Building a Prototype Data Analysis as a Service : The Stfcexperience." In *NOBUGS 2016 Proceedings* -

*New Opportunities for Better User Group Software*, edited by Tobias Richter, 23–28. ESS. https://doi.org/10.17199/NOBUGS2016.65.

Guru, Siddeswara, Ivan C. Hanigan, Hoang Anh Nguyen, Emma Burns, John Stein, Wade Blanchard, David Lindenmayer, and Tim Clancy. 2016. "Development of a Cloud-Based Platform for Reproducible Science: A Case Study of an Iucn Red List of Ecosystems Assessment." *Ecological Informatics* 36: 221–30. https://doi.org/https://doi.org/10.1016/j.ecoinf.2016.08.003.

Jones, Richard W. M. 2011. *Virt-Customize*. https://www.libguestfs.org/virt-customize.1.html.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. "Jupyter Notebooks ? A Publishing Format for Reproducible Computational Workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by Fernando Loizides and Birgit Scmidt, 87–90. IOS Press. https://eprints.soton.ac.uk/403913/.

NoMachine. 2023. "NX/Nomachine Remote Access for Everybody." https://www.nomachine.com/.

Picca, Frédéric, and Emmanuel Farhi. 2022. *SOLEIL Infra-Config*. https://gitlab.com/soleil-data-treatment/infra-config.

Randles, Bernadette M., Irene V. Pasquetto, Milena S. Golshan, and Christine L. Borgman. 2017. "Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study." In *2017 Acm/Ieee Joint Conference on Digital Libraries (Jcdl)*, 1–2. https://doi.org/10.1109/JCDL.2017.7991618.

StarNet Com. Santa Clara CA, USA. 2023. "StarNet Fastx Remote Linux X Windows." https://www.starnet.com/fastx/.

Tejs Vegge, et al. 2020. *BIG-Map. EU H2020 Grant Agreement No 957189*. https://doi.org/10.3030/957189.

Wang, Chunpeng, Ullrich Steiner, and Alessandro Sepe. 2018. "Synchrotron Big Data Science." *Small* 14 (46): 1802291. https://doi.org/10.1002/smll.201802291.