UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Containerized Jupyter Notebooks

Balancing flexibility and performance

Francesco Faenza, Emiliano Maccaferri, Claudia Canali

Cloud Edge and Security research group

University of Modena and Reggio Emilia

October 25, 2024

TOC

Introduction

Research question: How can an extensible, comprehensive, and secure environment for collaborative research be designed to provide transparent access to containerized resources? What are the performance cost of moving to a containerized environment?

Research Methodology: Software development and performance test.

Authors: Francesco Faenza, Emiliano Maccaferri, and Claudia Canali from the University of Modena and Reggio Emilia, Italy.

Related works

NaaVRE: Z. Zhao, S. Koulouzis, R. Bianchi, S. Farshidi, Z. Shi, R. Xin, Y. Wang, N. Li, Y. Shi, J. Timmermans *et al.*, "Notebook-as-a-VRE (NaaVRE): From private notebooks to a collaborative cloud virtual research environment," *Software: Practice and Experience*, vol. 52, no. 9, pp. 1947–1966, 2022.

Agave Platform:R. Dooley, S. R. Brandt, and J. Fonner, "The Agave Platform: An open, science-as-a-service platform for digital science," in *Proceedings of the Practice and Experience on Advanced Research Computing*, pp. 1–8, 2018.

Jupyter ESS: L. Fernández, R. Andersson, H. Hagenrud, T. Korhonen, E. Laface, B. Zupanc, *et al.*, "Jupyterhub at the ESS. An interactive Python computing environment for scientists and engineers," in *Proceedings of the 7th International Particle Accelerator Conference (IPAC 2016)*, 2016.

Japper: I. L. Kim, L. Zhao, C. Song, W. S. Neo, and B. Kelleher, "Japper: A comprehensive framework for streamlining Jupyter-based scientific web application development," in *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, pp. 1–4, 2024.

Where we started & where we are

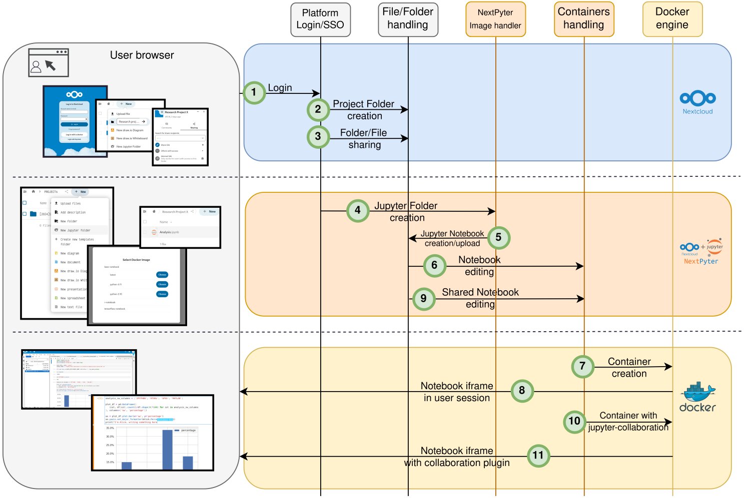
- **2019**: Started PhD and noticed the need for easier handling of Jupyter remote authentication
- **2022**: Published short paper in PEARC22 presenting an idea of integrating Jupyter with a spawner working also as auth system
- **2023**: Started from feedback received at PEARC24 isolating spawner from NextCloud
- **2024**: Integrated K8s working on other integrations started benchmarking

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

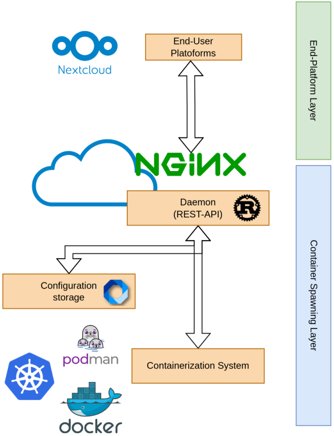
NextPyter Architecture

Technology Stack:

- Vue.js frontend
- PHP backend
- REST-API calls to NextPyter



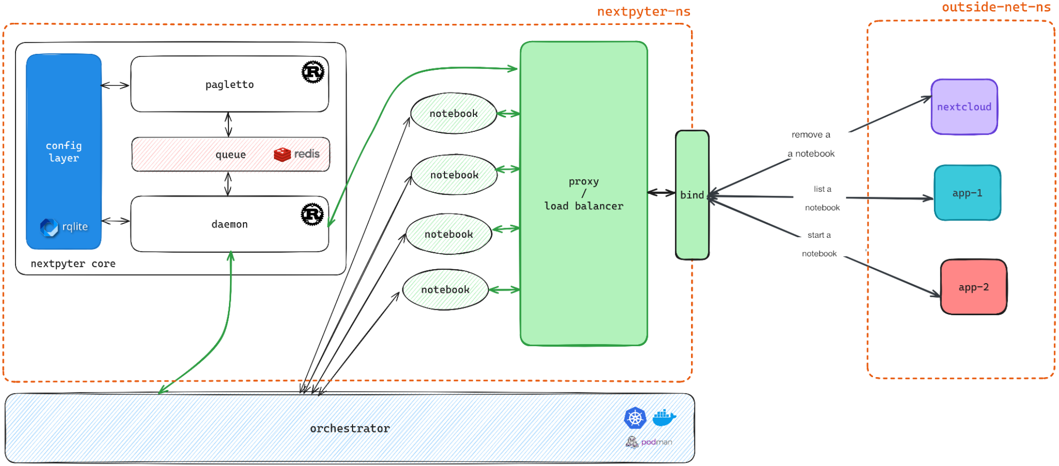
New architecture



Features:

- Move the NextPyter platform towards a microservices architecture with scalable design to enhance flexibility.
- Develop a core component to interface with container orchestration systems (Docker/Kubernetes) to be managed via a unified REST API.
- Insert a reverse proxy (Nginx) for container routing and incorporates a distributed configuration layer (RQLite) for image management.
- Insert an authentication layer (Keycloak) to provide a flexible, robust solution for both Docker and Kubernetes environments, supporting hybrid login systems.

Detailed Architecture



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Current Stage

Where We Are Now:

A **K8s** based instance is being used by our internal NextCloud and being tested by part of our group.

Lot of developer worked on the project recently so we are **uniforming** the base code and creating **documentation**.

The repositories hosting the code for the needed services: `https://gitlab.com/nextpyter`

Performance evaluation

A performance test was conducted to evaluate the overhead of **containerized Jupyter** Notebooks compared to bare-metal implementations

The test focused on common **data science scenarios**: local execution of Python scripts vs. Jupyter Notebook execution

Previous research suggests that performance differences between Jupyter Notebooks and Python scripts are minimal, with only a slight overhead in CPU load for notebooks

P. Prathanrat and C. Polprasert, "Performance prediction of Jupyter notebook in JupyterHub using machine learning," in *Proceedings of the 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, vol. 3, pp. 157–162, 2018.

K. Grotov, S. Titov, V. Sotnikov, Y. Golubev, and T. Bryksin, "A large-scale comparison of Python code in Jupyter notebooks and scripts," in *Proceedings of the 19th International Conference on Mining Software Repositories*, pp. 353–364, 2022.

Testing environment

Pyperfomence used, analysys focused on data anlysis related benchmarks: bare metal, docker, jupyter on bare metal, and jupyter on docker

Tests were conducted on a Proxmox cluster and a couple of Raspberry Pi (repeated multiple times).

Performance metrics:

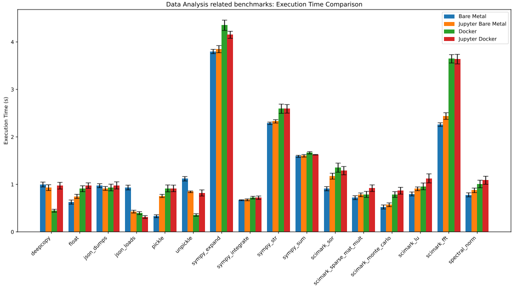
- execution time
- memory load
- CPU load

Metrics were normalized using Z-scores and reliability checked with ANOVA (Analysis of Variance) and Tukey's HSD (Honestly Significant Difference)

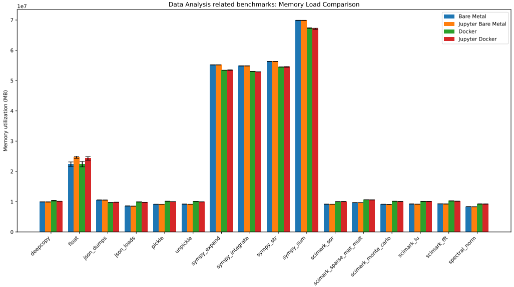
Execution time and Memory load

No statistically significant difference in Execution time and Memory load

Execution time

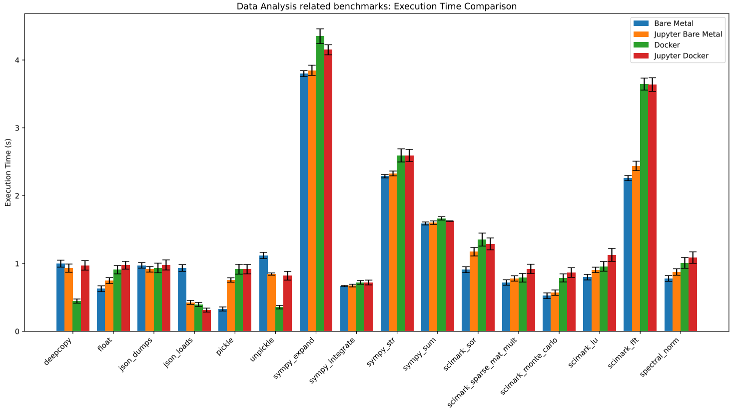


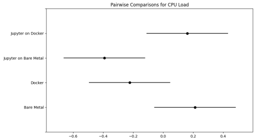
Memory load



CPU load

Execution time





Statistical Significant differences:

- bare metal vs Jupyter on Bare Metal
- Jupyter on Bare Metal vs Jupyter Dockerized

Advantages and Limitations

Advantages

- Integration of Notebooks within any web application
- Real-time collaboration capabilities on Notebooks
- Abstraction of technical complexities
- Process isolation granted by Docker/Podman/K8s architecture
- Multi-host configuration possibility
- Self-hosting advantages (sensitive data privacy among others)

Limitations

- Missing Object-storage driver for Docker/Podman
- Persistence of personalized environments (working on it)
- Limiting use to specific users and fine-tuning of resources (working on it)

Future Work

Federation: Looking into federation protocols and way to intgrate them in the architecture

Consuption feedback: Integrating a way to provide within API domain a way to obtain performance of a spceific notebook/cell execution

Authors: Francesco Faenza, Emiliano Maccaferri, and Claudia Canali from the University of Modena and Reggio Emilia, Italy.

Any questions?

Personal contacts

Francesco Faenza

francesco.faenza@unimore.it

Security, Edge and Cloud Lab

https://secloud.ing.unimore.it/people/faenza/

”Never doubt that a small group of thoughtful, committed citizens can change the world; indeed, it is the only thing that ever has.”

– Margaret Mead

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

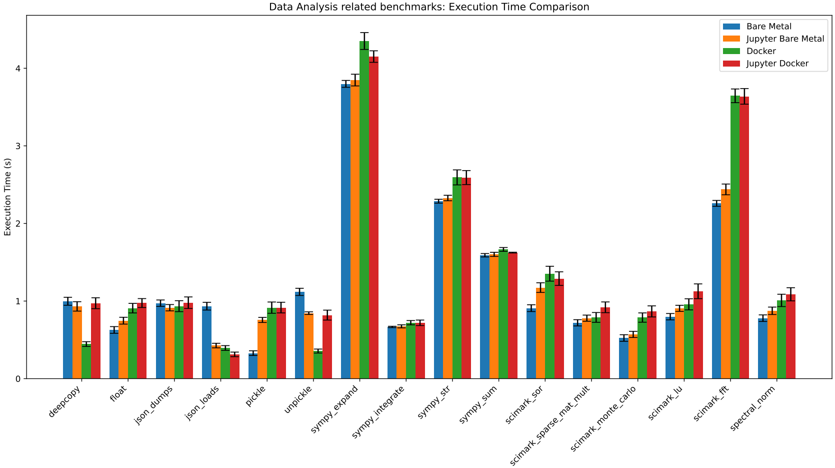
Anova and Tucky result on CPU

ANOVA (F-statistic = 0.0223, p-value = 0.9954)

Tukey Test Results for Normalized CPU Load:

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
Bare Metal	Docker	-0.4351	0.154	-0.9737	0.1035	False
Bare Metal	Jupyter on Bare Metal	-0.603	0.0223	-1.1416	-0.0644	True
Bare Metal	Jupyter on Docker	-0.0579	0.9919	-0.5965	0.4808	False
Docker	Jupyter on Bare Metal	-0.1679	0.8431	-0.7065	0.3707	False
Docker	Jupyter on Docker	0.3773	0.2602	-0.1614	0.9159	False
Jupyter on Bare Metal	Jupyter on Docker	0.5451	0.0462	0.0065	1.0838	True

CPU load



Execution time

Detailed K8s Architecture