

# Machine Learning Performance Engineer

## Description

We are looking for an engineer with experience in low-level systems programming and optimisation to join our growing ML team.

Machine learning is a critical pillar of Jane Street's global business. Our ever-evolving trading environment serves as a unique, rapid-feedback platform for ML experimentation, allowing us to incorporate new ideas with relatively little friction.

Your part here is optimising the performance of our models – both training and inference. We care about efficient large-scale training, low-latency inference in real-time systems and high-throughput inference in research. Part of this is improving straightforward CUDA, but the interesting part needs a whole-systems approach, including storage systems, networking and host- and GPU-level considerations. Zooming in, we also want to ensure our platform makes sense even at the lowest level – is all that throughput actually goodput? Does loading that vector from the L2 cache really take that long?

If you've never thought about a career in finance, you're in good company. Many of us were in the same position before working here. If you have a curious mind and a passion for solving interesting problems, we have a feeling you'll fit right in.

There's no fixed set of skills, but here are some of the things we're looking for:

- An understanding of modern ML techniques and toolsets
- The experience and systems knowledge required to debug a training run's performance end to end
- Low-level GPU knowledge of PTX, SASS, warps, cooperative groups, Tensor Cores and the memory hierarchy
- Debugging and optimisation experience using tools like CUDA GDB, NSight Systems, NSight Computesight-systems and nsight-compute
- Library knowledge of Triton, CUTLASS, CUB, Thrust, cuDNN and cuBLAS
- Intuition about the latency and throughput characteristics of CUDA graph launch, tensor core arithmetic, warp-level synchronization and asynchronous memory loads
- Background in Infiniband, RoCE, GPUDirect, PXN, rail optimisation and NVLink, and how to use these networking technologies to link up GPU clusters
- An understanding of the collective algorithms supporting distributed GPU training in NCCL or MPI
- An inventive approach and the willingness to ask hard questions about whether we're taking the right approaches and using the right tools
- Fluency in English

*If you're a recruiting agency and want to partner with us, please reach out to [agency-partnerships@janestreet.com](mailto:agency-partnerships@janestreet.com).*

## How the process will look like

Your teammates will gather all requirements within our organization. Then, once priority has been discussed, you will decide as a team on the best solutions and architecture to meet these needs. In continuous increments and continuous

## Hiring organization

Candidate-1st

## Employment Type

Full-time

## Beginning of employment

asap

## Job Location

London, England, United Kingdom

## Working Hours

40

## Base Salary

euro GBP 40K - 74K \*

## Date posted

May 21, 2024

communication between the team and stakeholders, you're part of making data play an even more important (and understood) part withing Brand New Day.

### **Job Benefits**

GBP 40K – 74K \*