About

UPMEM is a leading European innovative fabless semiconductor startup with a recognized global leadership in “Processing in Memory” (PIM). UPMEM develops and ships a game changing compute acceleration solution which massively reduces cost and energy footprint in the datacenter, solving the Memory wall. UPMEM has top tier customers in US and China.

Founded in 2015, the company relies on an expert, entrepreneurial and international 25 people team mainly based in Grenoble.

UPMEM has leading industrial and financial backing from Europe, America and Asia.

UPMEM was recently selected among the top 100 innovative Silicon 100 startups by EETimes and was awarded the EU H2020 EIC challenge backing for deep tech...

Job Description

Contribute to the definition of AI PIM platform for LLMs, develop AI PIM back-end for de-facto standard AI frameworks.

What you’ll be doing:

- Understand, analyze, profile, and optimize LLM workloads for AI PIM platform
- Collaborate with researchers and engineers in UPMEM and externally, providing insights useful for HW design and work on improving of the performance of third party workloads
- Design and implement production-quality AI framework backend
- Understand and solve challenges pertaining to a hybrid AI PIM platform

Preferred Experience

- Experience with enabling and optimization of training or/and inference of specific LLMs with more than 100B parameters on target HW
- Familiarity with Deep Learning and Transformers/LLMs from open source repos like Hugging Face Hub
- Experience with modern AI frameworks (Tensorflow, Pytorch) - 3+ years
- Experience with Python and C; 5+ years
- Familiarity with DeepSpeed, Horovod or other distributed training frameworks would be a plus

Recruitment Process

Please send application and resume. Interviews with tech and management team.

Additional Information

- Contract Type: Full-Time
- Location: Grenoble, Paris
- Education Level: Master's Degree
- Experience: > 2 years
- Possible partial remote

APPLY NOW