KLA Overview:

KLA is a global leader in diversified electronics for the semiconductor manufacturing ecosystem. Virtually every electronic device in the world is produced using our technologies. No laptop, smartphone, wearable device, voice-controlled gadget, flexible screen, VR device or smart car would have made it into your hands without us. KLA invents systems and solutions for the manufacturing of wafers and reticles, integrated circuits, packaging, printed circuit boards and flat panel displays. The innovative ideas and devices that are advancing humanity all begin with inspiration, research and development. KLA focuses more than average on innovation and in 2019 we invested 15% of sales back into R&D. Our expert teams of physicists, engineers, data scientists and problem-solvers work together with the world's leading technology providers to accelerate the delivery of tomorrow's electronic devices. Life here is exciting and our teams thrive on tackling really hard problems. There is never a dull moment with us.

Role Description

- Building theoretical models that break down KLA's image processing algorithms, that leverage AI, in computing terms such as bandwidth, computational FLOPS, etc.
- Bridging the gap between the theoretical peak performance achievable on current and next-gen hardware such as GPUs and AI accelerators by enhancing the algorithm.
- Porting and optimizing algorithms on current and next-gen CPUs, GPUs, and AI accelerators by leveraging constructs in high-performance modern programming languages such as C++-14/C++-17, and low-level programming models such as SIMD extensions (SSE/AVX), CUDA, OpenVINO, etc.
- Exploring paths to achieve price-optimized-performance in next-generation devices that implement revolutionary new solutions to accelerate AI algorithms for training and inference.

Expected Background

- New/recent college graduates in Ph.D, MS in EE/CS/CSE. Bachelors graduates will also be considered with exceptional background and prior experience in HPC field.
- Strong foundation in computer architecture, with interest in high performance parallel processing at the device level (GPUs or CPUs/SIMD).
- Strong mental model of computational loads and mapping different algorithms to parallel architectures.
- Proficient in programming skills in C/C++/Python.
- Good understanding and exposure to the Linux operating system at the user level.
- Exposure to multiprocessor and multithreading concepts
- A self-motivated individual with good communication skills.

Bonus Skills

- Hands-on experience with GPU programming using CUDA, OpenCL or SYCL, and modern CPU programming constructs such as those in C++-14 / C++-17
- Exposure to profiling tools such as NSIGHT or VTUNE.
- Experience with large-scale distributed HPC systems.
- Familiarity with AI frameworks like TensorFlow.
- Hands-on work in developing and optimizing computer vision algorithms at scale.

We offer a competitive, family friendly total rewards package. We design our programs to reflect our commitment to an inclusive environment, while ensuring we provide benefits that meet the diverse needs of our employees.

KLA is proud to be an equal opportunity employer