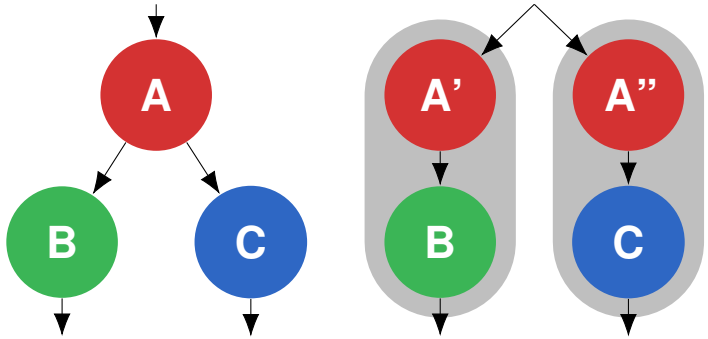
Kernel Fusion by Duplication on Graphics Processing Units

Applications in image processing typically consist of multiple kernels. Data dependencies between those kernels usually manifest in slow memory accesses, especially when targeting highly parallel accelerators such as graphics processing units (GPUs). One way to reduce the number of such costly accesses is to combine multiple kernel bodies so that some slow memory accesses are replaced by faster ones. This technique is called *kernel fusion*.



Whether the fusion of a specific set of kernels is beneficial depends on multiple factors. In this thesis, the fusion of a specific class of dependency patterns shall be investigated, which requires the duplication of some kernel bodies, trading an increase in computational effort for an improvement of locality, and therefore a potential gain in overall application performance. The aims of this thesis are a) Understanding common performance bottlenecks in GPU applications and internalizing the motivation of related optimization techniques such as kernel fusion, b) Getting acquainted with the Hipacc image processing framework and investigating the performance impact when fusing certain application patterns, and c) Implementing an automatic way of fusing the mentioned application patterns in Hipacc.

Required skills:	Good knowledge of C++, basic knowledge of CUDA may be helpful
Nature of work:	Theory (30 %), Conception (10 %), Implementation (60 %)
Contact:	Marvin Lunz (marvin.lunz@fau.de).

Lehrstuhl für Informatik 12
Hardware-Software-Co-Design
Cauerstraße 11
91058 Erlangen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT