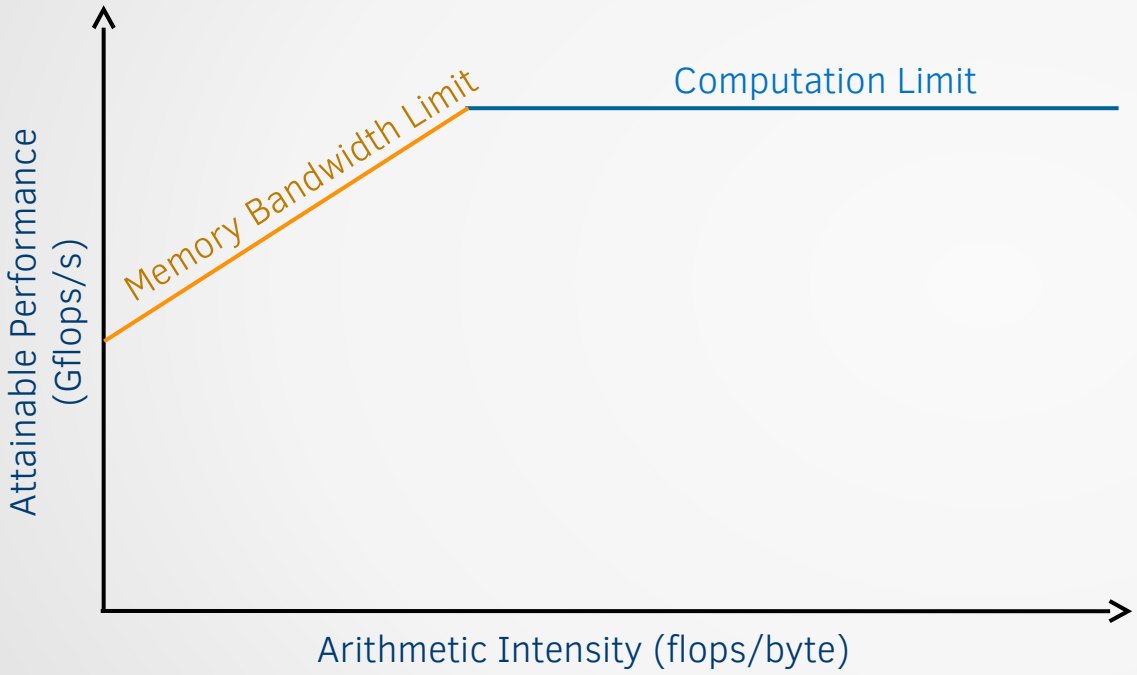
Roofline analysis

Mathieu Lobet, Matthieu Haefele

Maison de la Simulation, CEA, CNRS, Université Paris-Sud, UVSQ,
Universite Paris-Saclay, F-91191 Gif-sur-Yvette, France
(mathieu.lobet@cea.fr)

Description of the roofline analysis

Roofline performance model



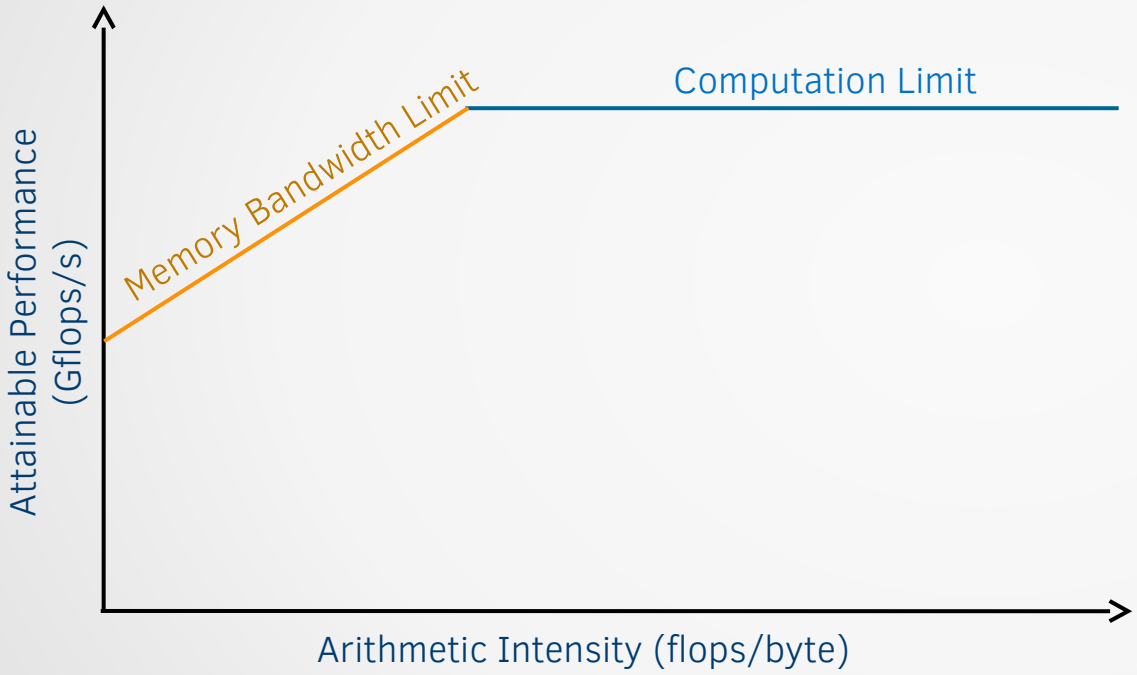
"Roofline is a visually intuitive performance model used to bound the performance of various numerical methods and operations running on multicore, manycore, or accelerator processor architectures."

Roofline:

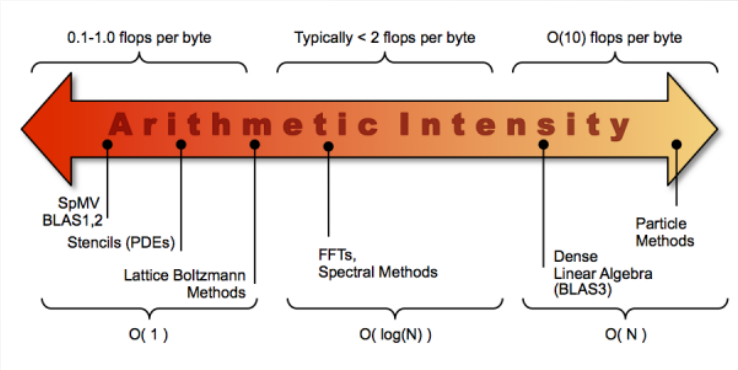
Reflects a performance bound (Gflops/s) as a function of Arithmetic Intensity (AI).

[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

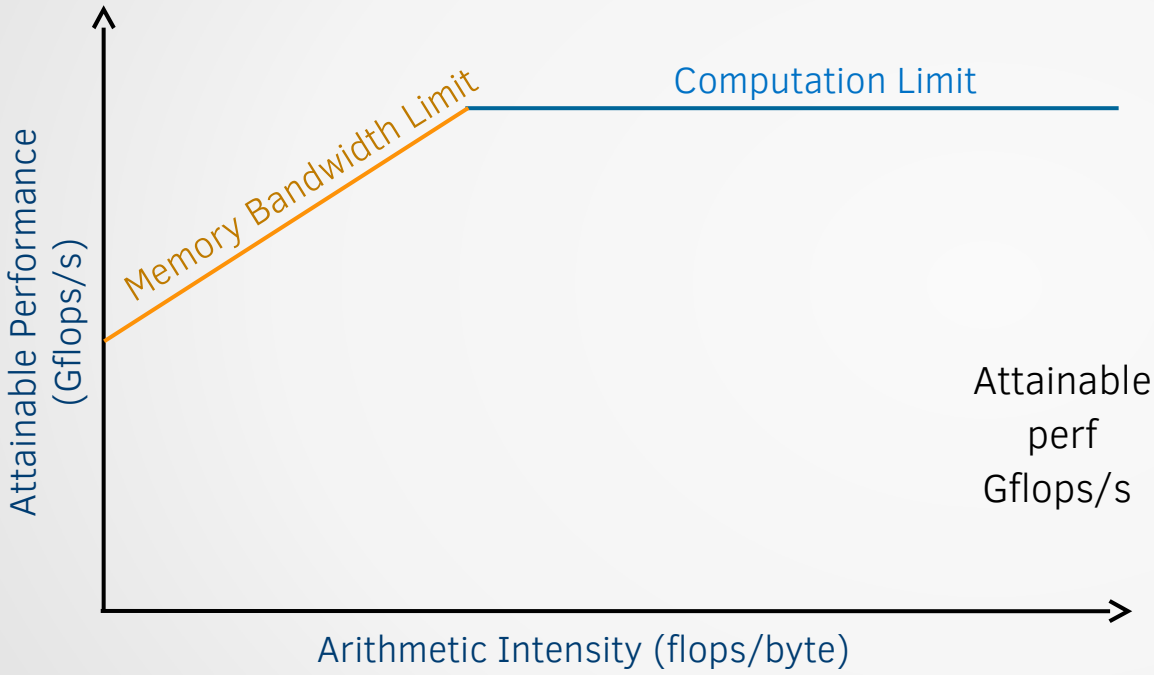


$$\text{Arithmetic intensity} = \frac{\text{Total flops computed}}{\text{Total bytes transfered from DRAM/cache}}$$



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

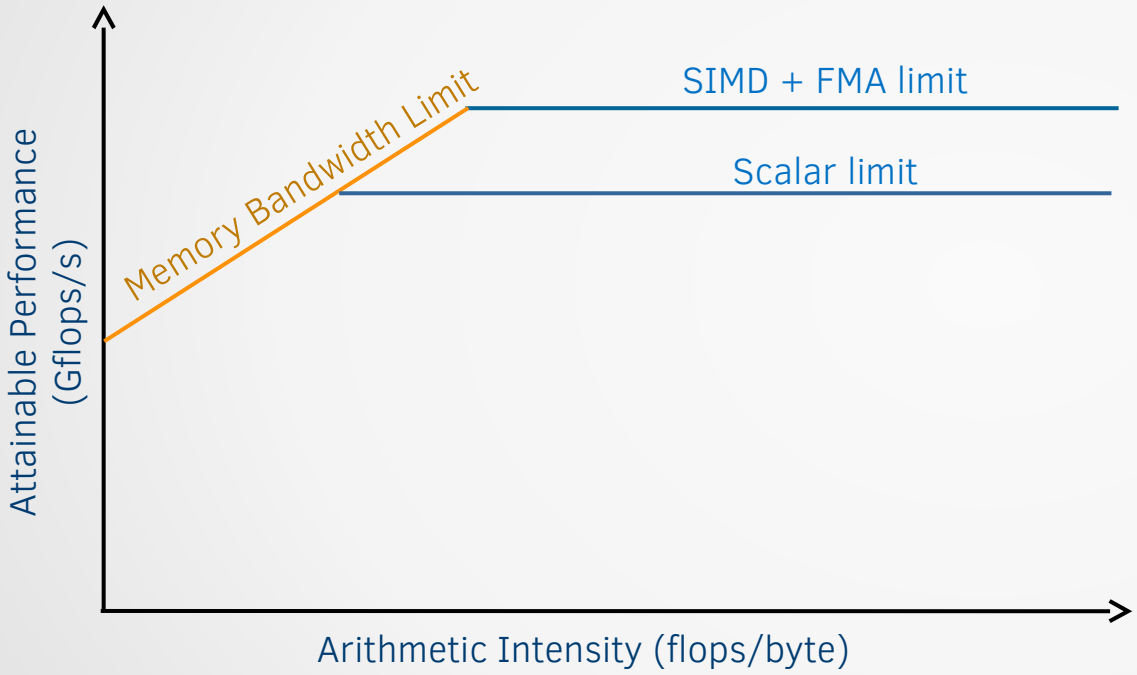


The attainable system performance is the maximal performance that can be reached by an application:

Attainable perf Gflops/s = min { Peak performace Gflops/s ; Peak Memory Bandwidth x Arithmetic Intensity }

[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model



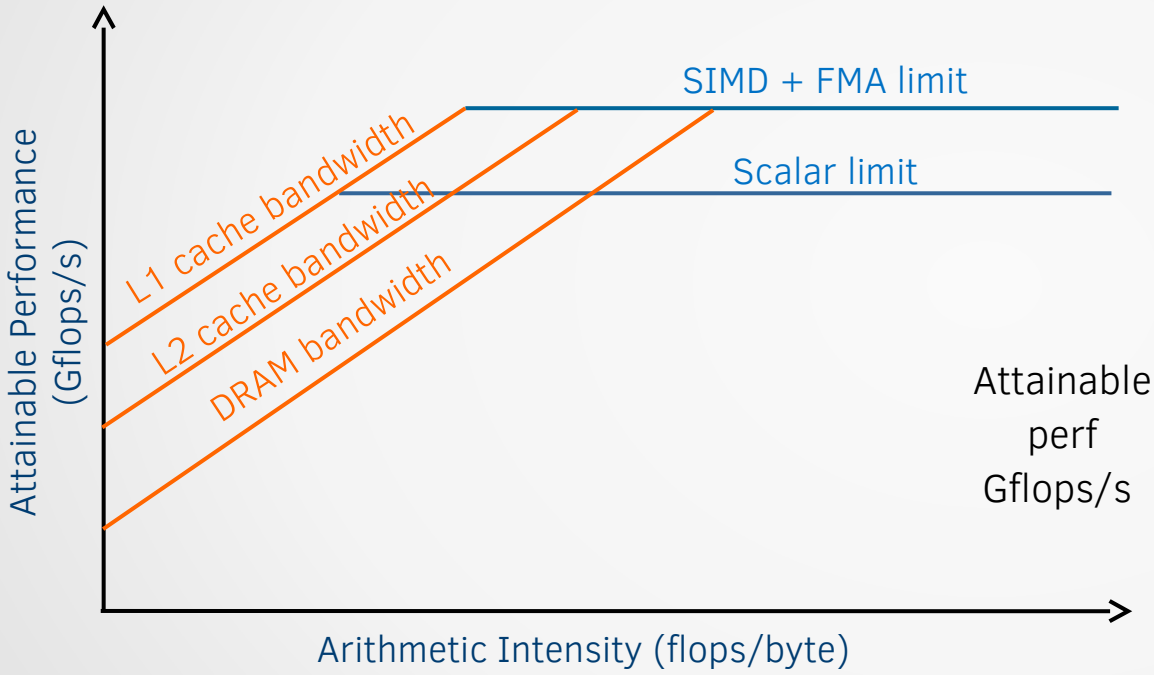
Peak performance = Frequency x vector width x vpus x FMA x number of Cores

Bandwidth can be taken as STREAM TRIAD value

Bandwidth and peak performance can be computed, e.g., using the Empirical roofline toolkit from LBNL [2]

[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

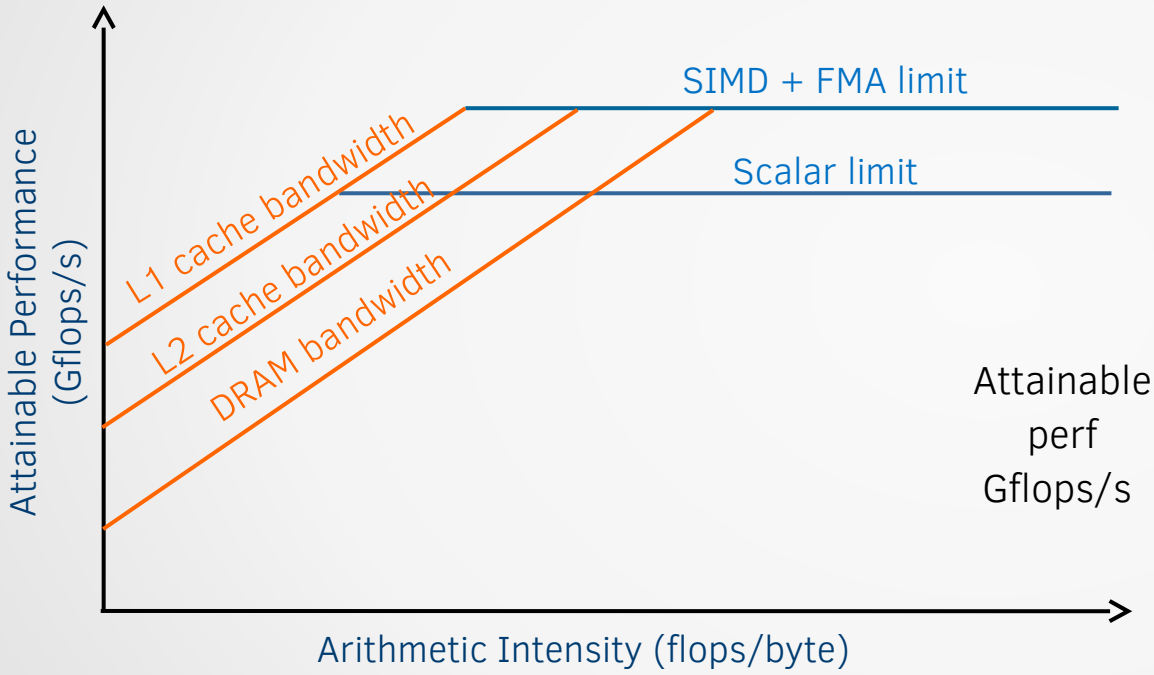


Total volume of bytes transferred across all memory hierarchies to the core:

Attainable perf Gflops/s = min { Peak performace Gflops/s ; Peak Memory Bandwidth x Arithmetic Intensity }

[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

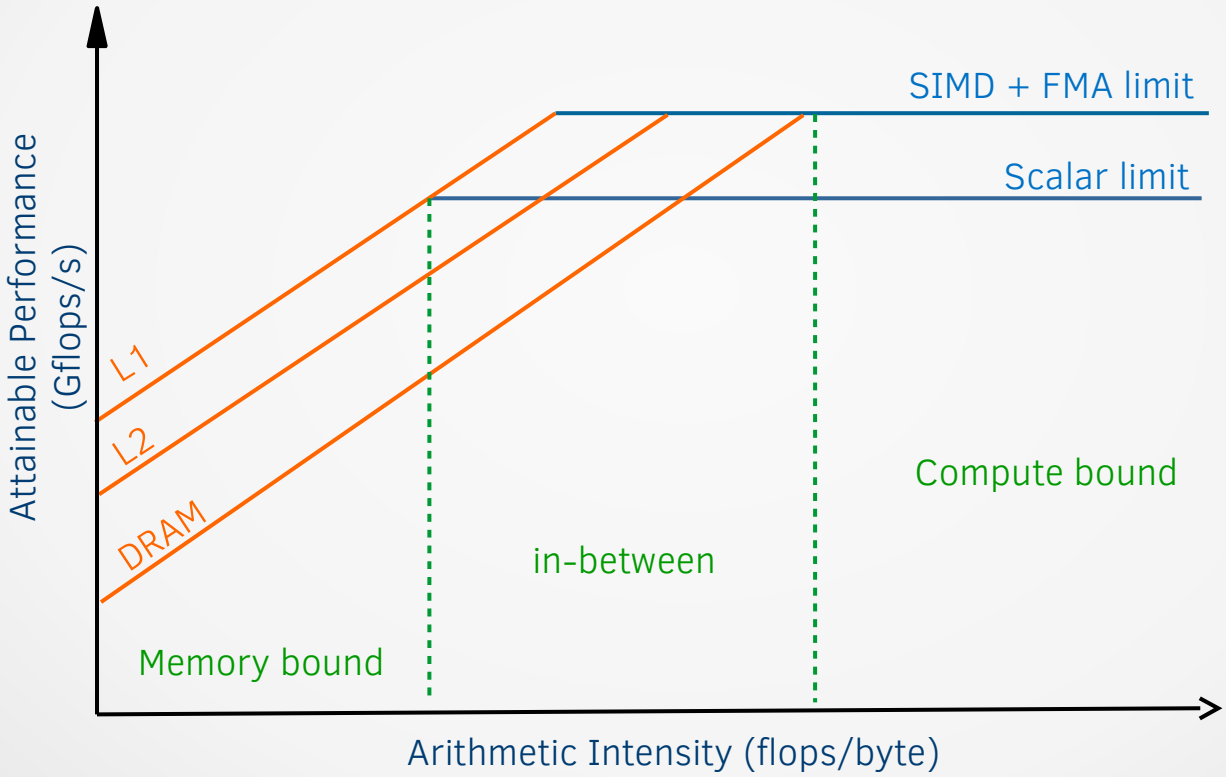


Total volume of bytes transferred across all memory hierarchies to the core:

Attainable perf Gflops/s = min { Peak performace Gflops/s ; Peak Memory Bandwidth x Arithmetic Intensity }

[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

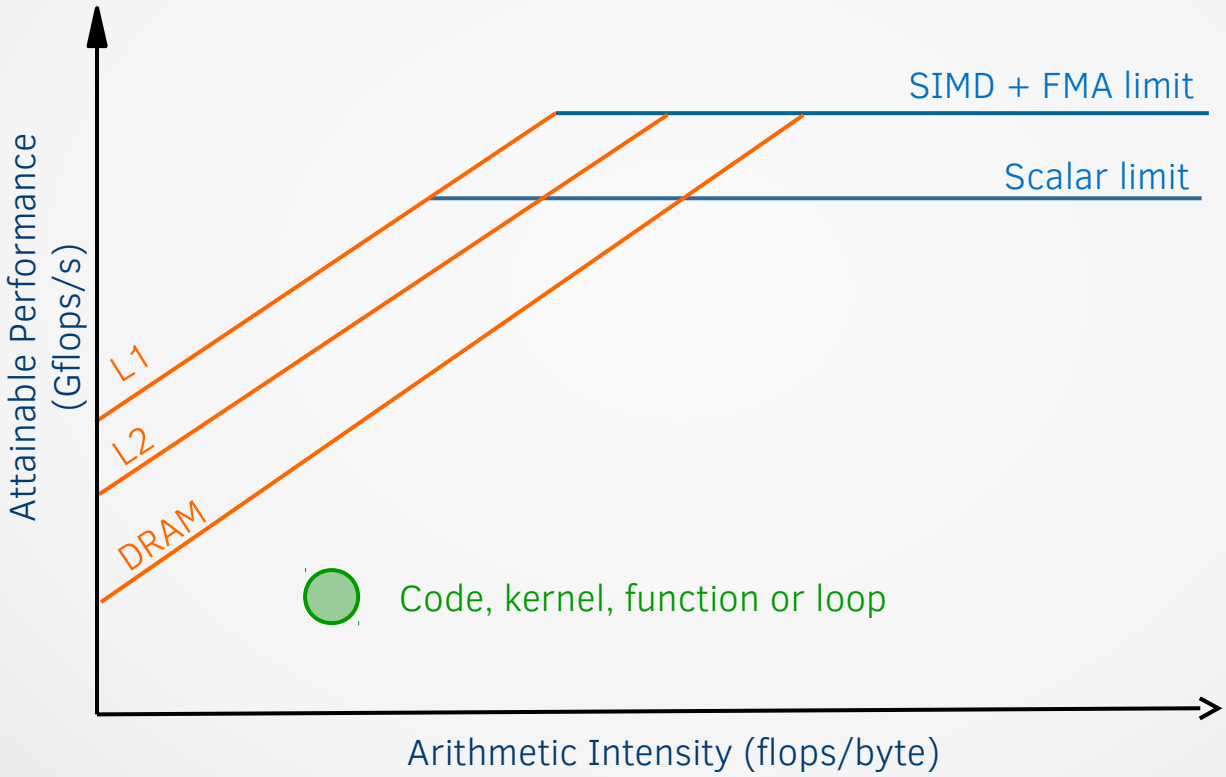
Roofline performance model



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

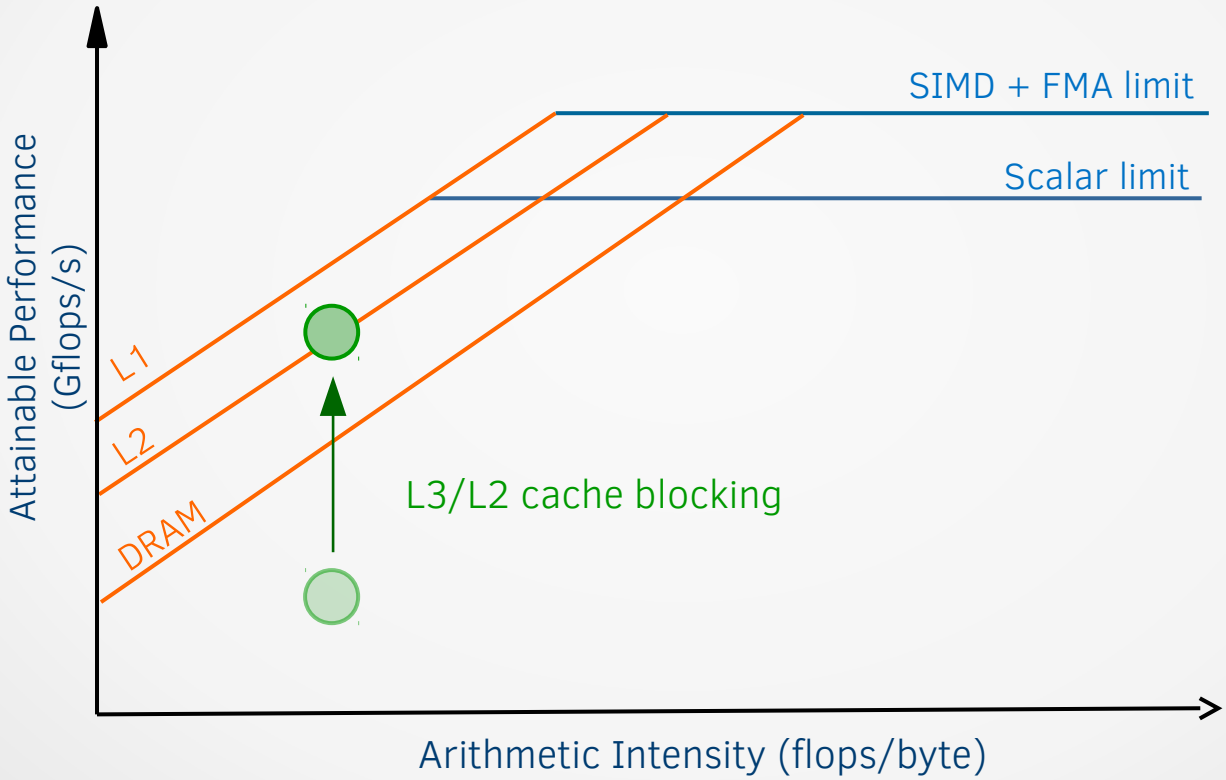
Roofline performance model

DRAM bound code



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

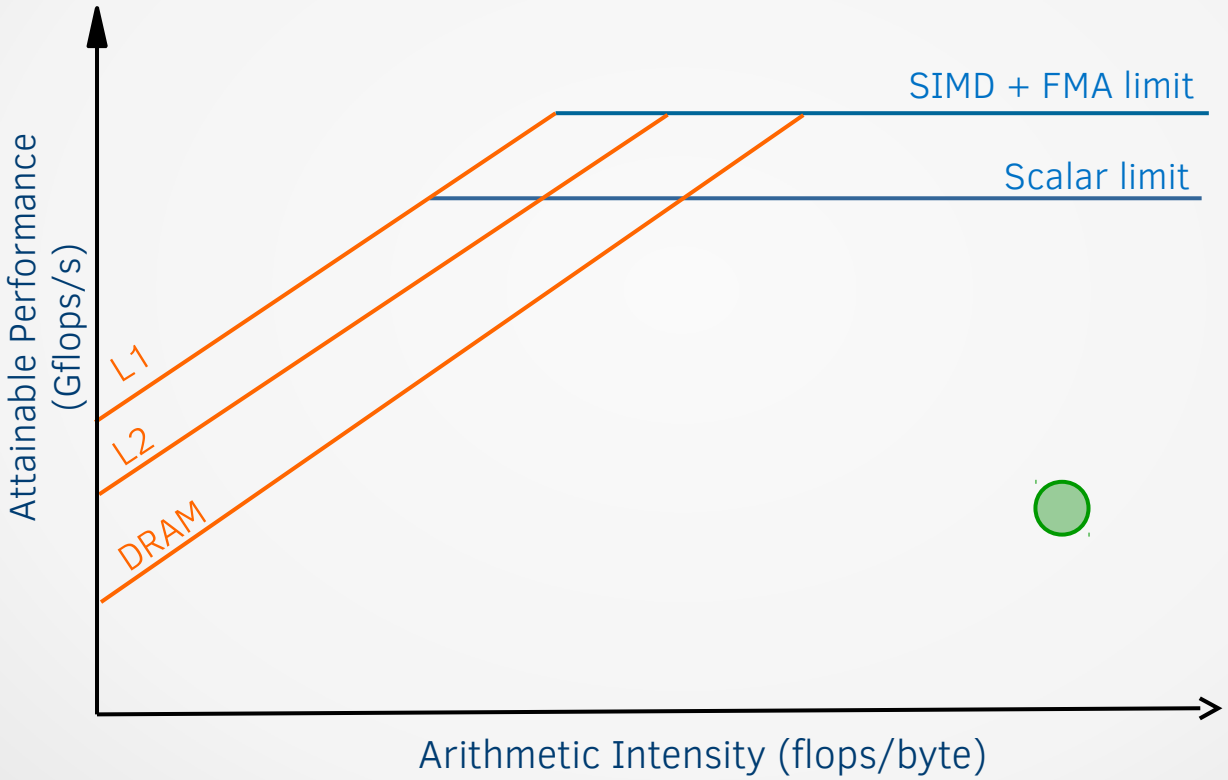
Roofline performance model



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

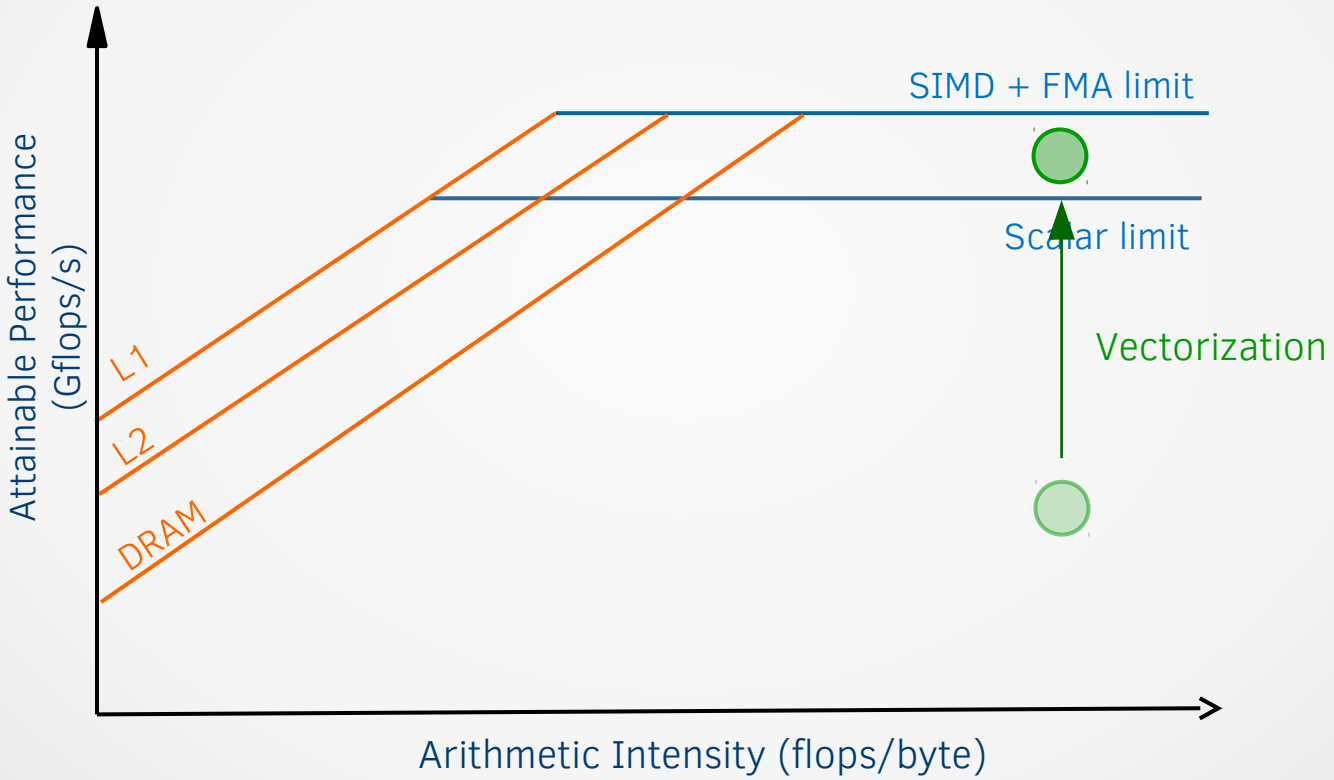
Compute bound code



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline

Roofline performance model

Compute bound code



[1] S. Williams et al. CACM (2009), crd.lbl.gov/departments/computer-science/PAR/research/roofline