

# Performance Portability for Next-Generation Heterogeneous Systems

Dr Tom Deakin

Lecturer in Advanced Computer Systems

University of Bristol

<b>Nov'23 Top500 Rank</b>	<b>System</b>	<b>Accelerator</b>
1	Frontier	✓
2	Aurora	✓
3	Eagle	✓
4	Supercomputer Fugaku	✗
5	LUMI	✓
6	Leonardo	✓
7	Summit	✓
8	MareNostrum 5 ACC	✓
9	Eos NVIDIA DGX SuperPOD	✓
10	Sierra	✓



Latency

Throughput

“Complex” cores

Instruction Level Parallelism

Deep cache hierarchy

NUMA

Wide SIMD

In-core accelerators

More “simple” cores

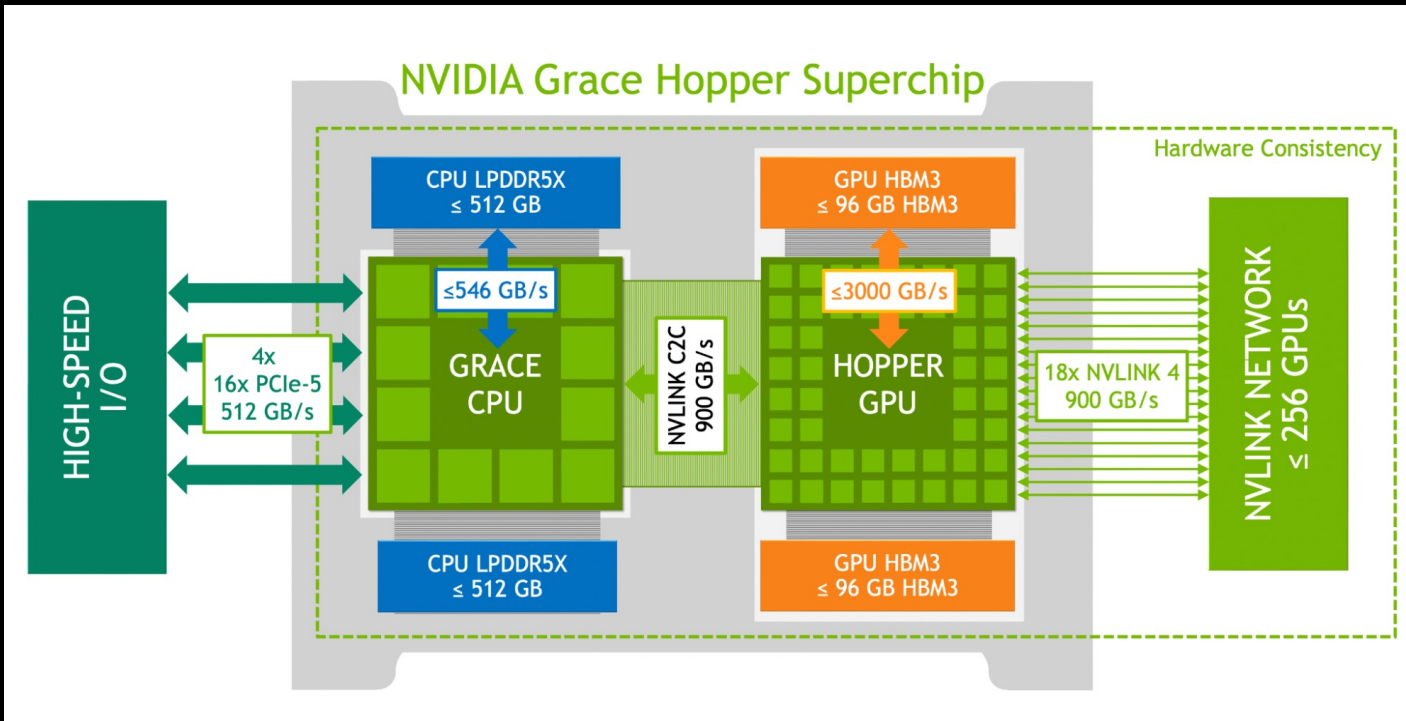
Very wide SIMD

Fast context switching

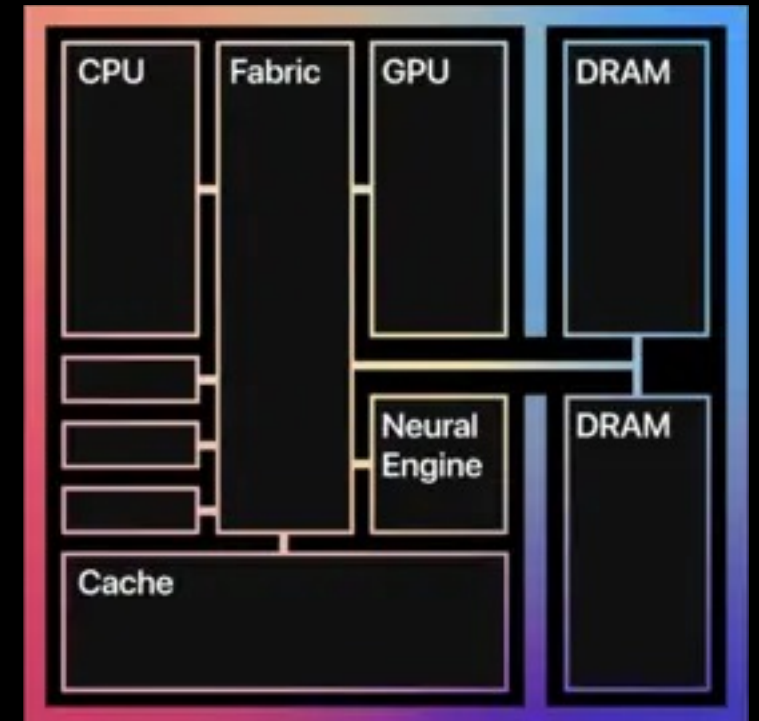
Programmable memory hierarchy

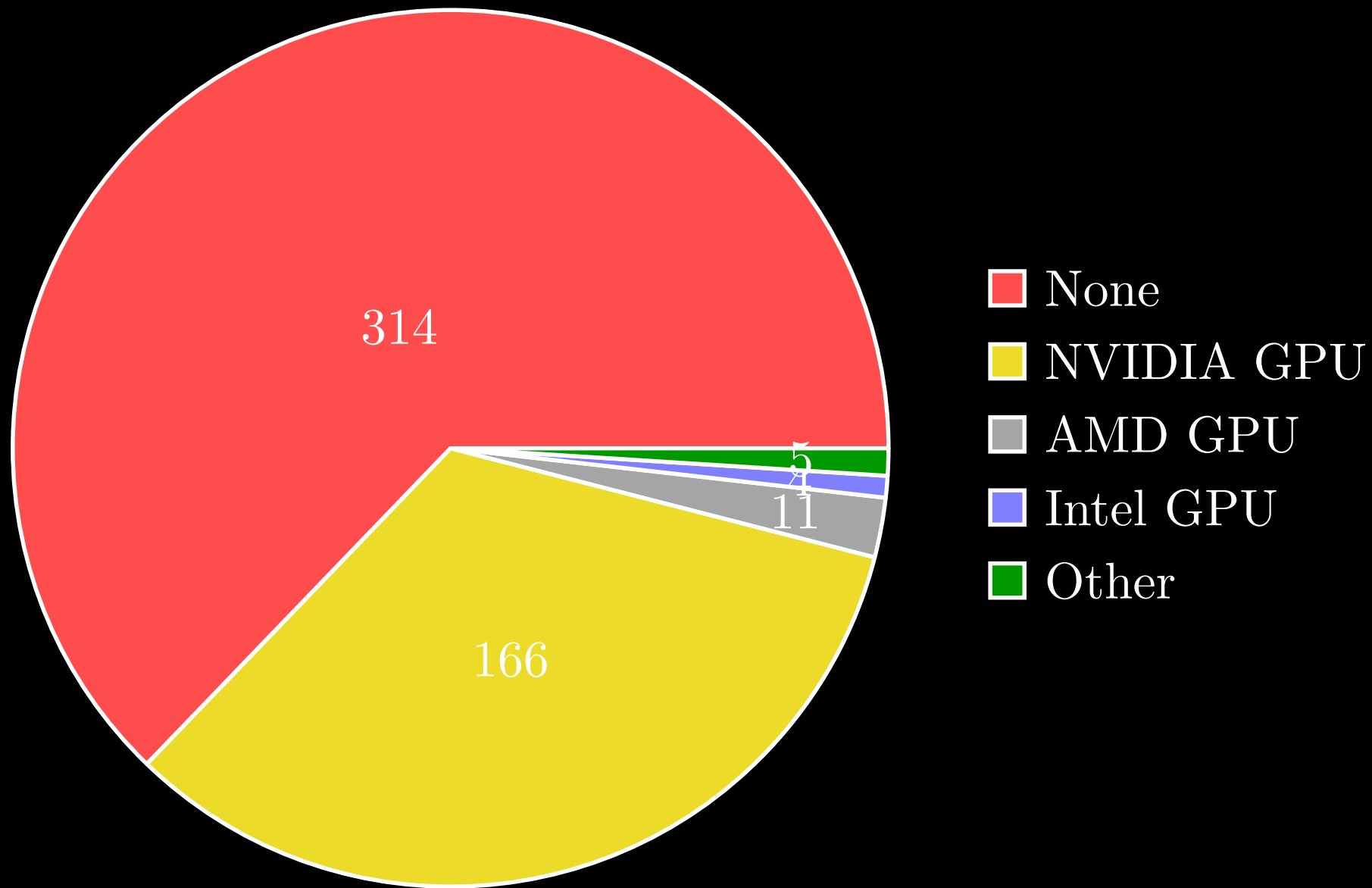
Latest memory technology

# NVIDIA Grace-Hopper



# Apple M1





Data: TOP500 November 2023

Updated version of chart from: [doi.org/10.1109/P3HPC56579.2022.00006](https://doi.org/10.1109/P3HPC56579.2022.00006)

Tension between migrating to next system  
(which may be GPUs), and keeping running  
on current system

# Performance, Portability, and Productivity

“A code is performance portable if it can achieve a similar fraction of peak hardware performance on a range of different target architectures”.



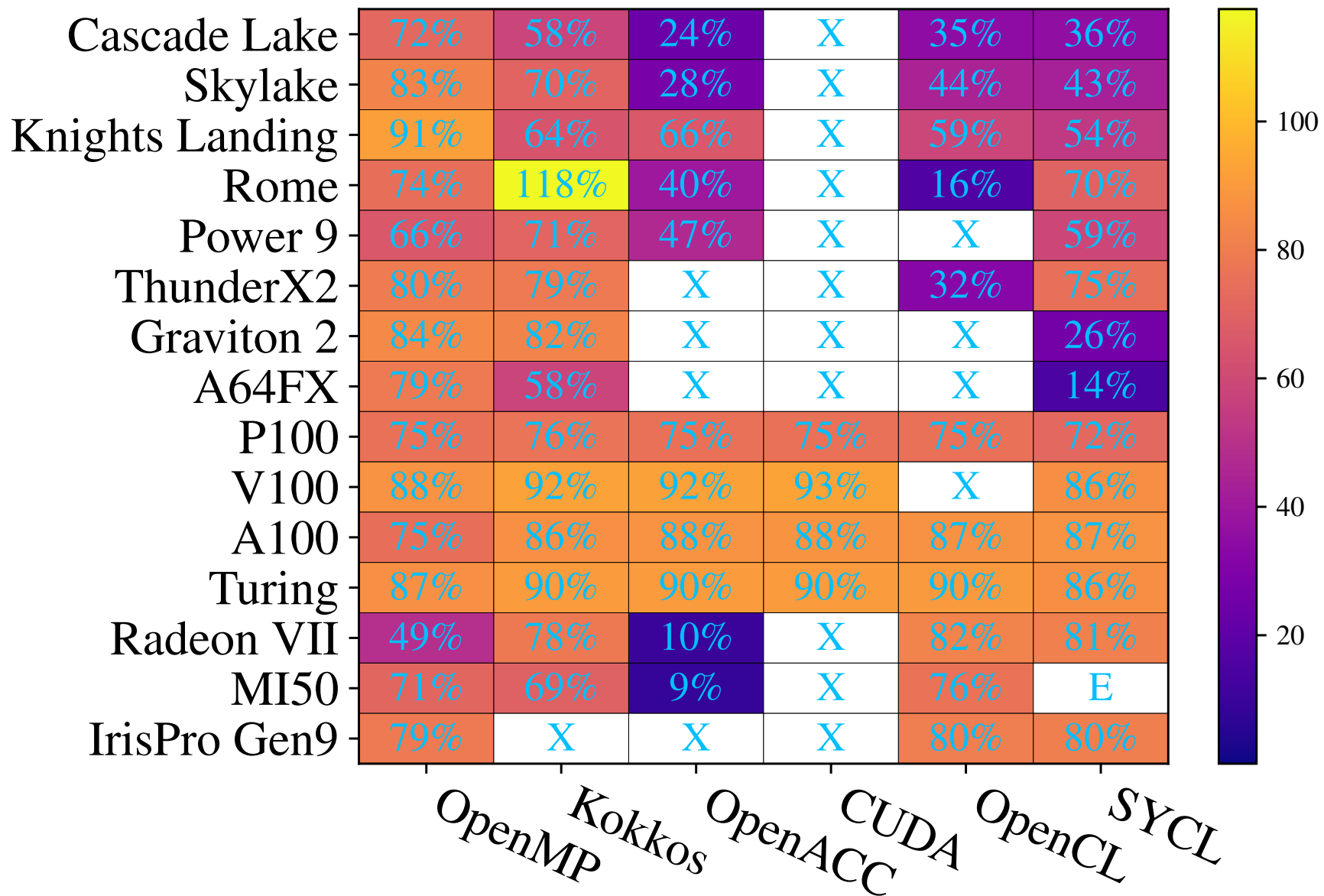
Problem

Application

Platform

Efficiency

BabelStream Triad array size=2\*\*25

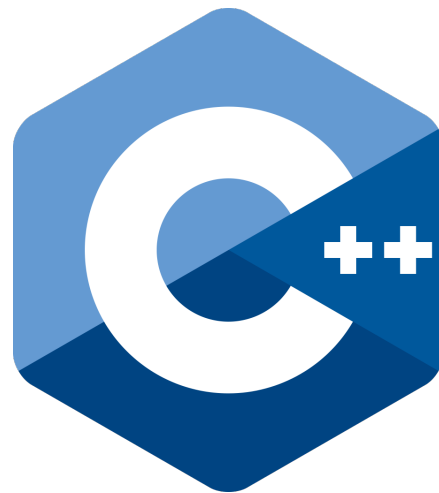


$\Phi$

Re  Frame

The ReFrame logo consists of a red square with a white outline, and a green square with a white outline, overlapping the red square.

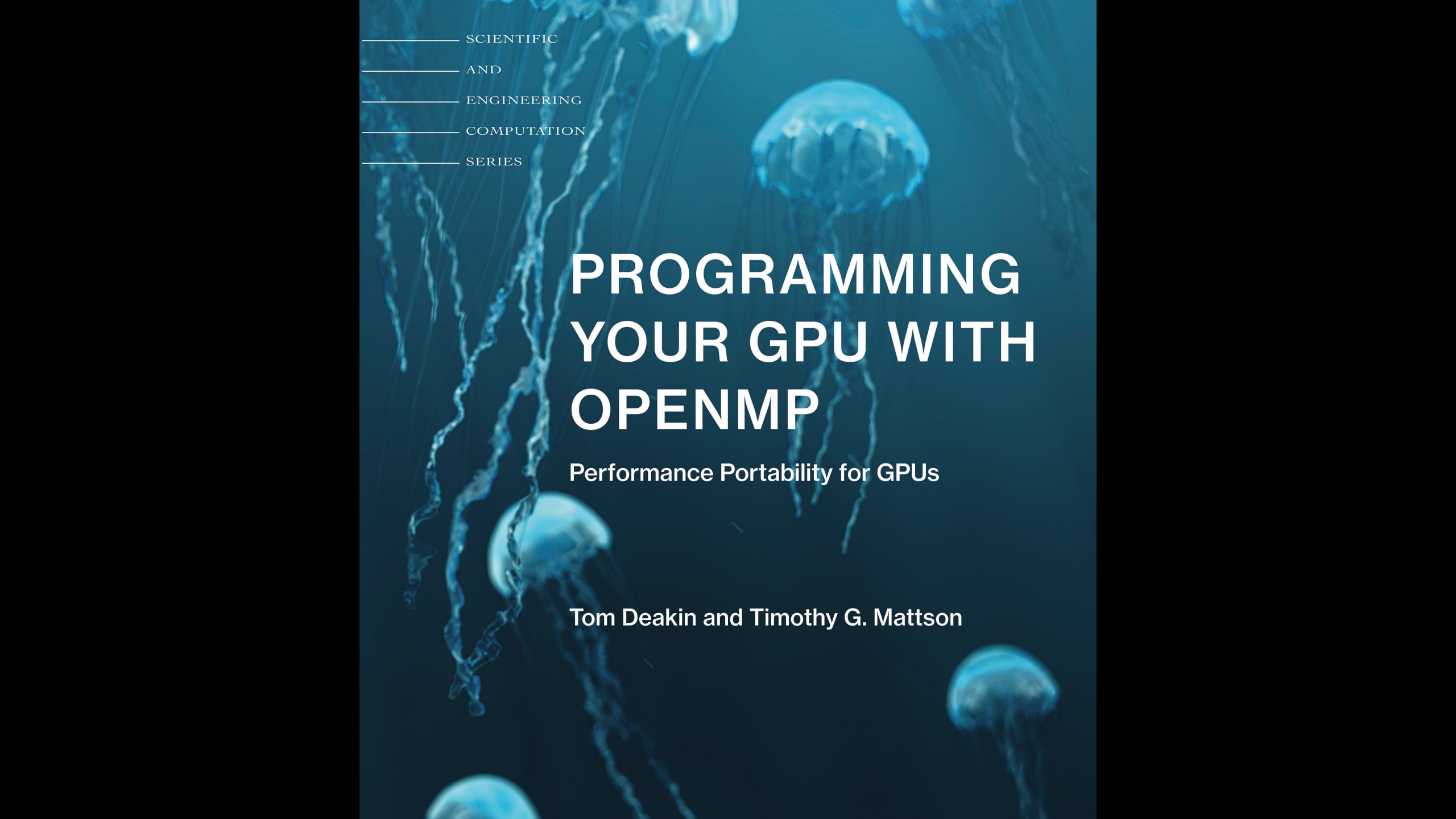
**Spack**





Prompt: A fight between parallel programming languages  
Generated with AI on Microsoft Bing Image Creator · 6 December 2023 at 11:58 am





SCIENTIFIC  
AND  
ENGINEERING  
COMPUTATION  
SERIES

# PROGRAMMING YOUR GPU WITH OPENMP

Performance Portability for GPUs

Tom Deakin and Timothy G. Mattson

Develop with P3 in mind with Standard Parallelism

Use open-standards as confluent off-ramp to be productive today

Express all concurrent work asynchronously

Build in tuning parameters

Test all compilers & runtimes, on all systems

Tell your vendor