# Computational Science on Many-Core Architectures

360.252

**Karl Rupp**

Institute for Microelectronics, TU Wien
http://www.iue.tuwien.ac.at/

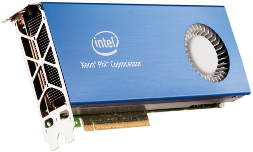
Zoom Channel 941 8518 8102
Q&A on Wednesday, October 12, 2022

# Introduction

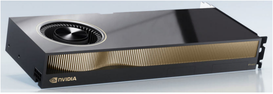
## Many-Core Architectures

- High FLOP/Watt ratio
- High memory bandwidth
- (Usually) Attached via PCI-Express

AMD W6800
17.8 TFLOPs FP32
512 GB/sec



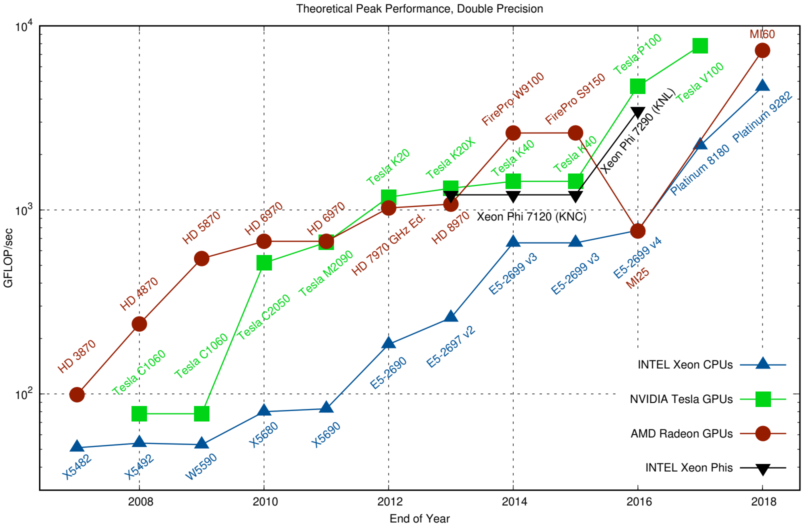
INTEL Xeon Phi
6.9 TFLOPs FP32
400+ GB/sec



NVIDIA RTX A6000
38.7 TFLOPs FP32
768 GB/sec

# Introduction

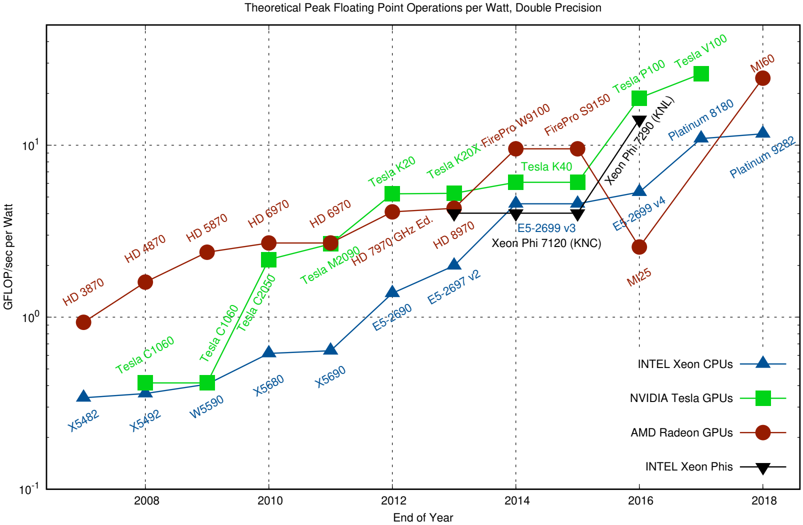
## Theoretical Peak Performance



https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

# Introduction

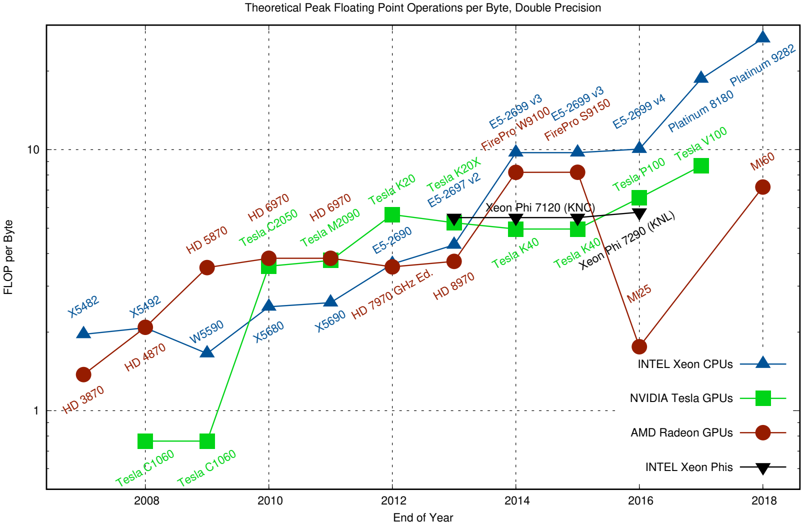
## Theoretical Peak Performance per Watt



https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

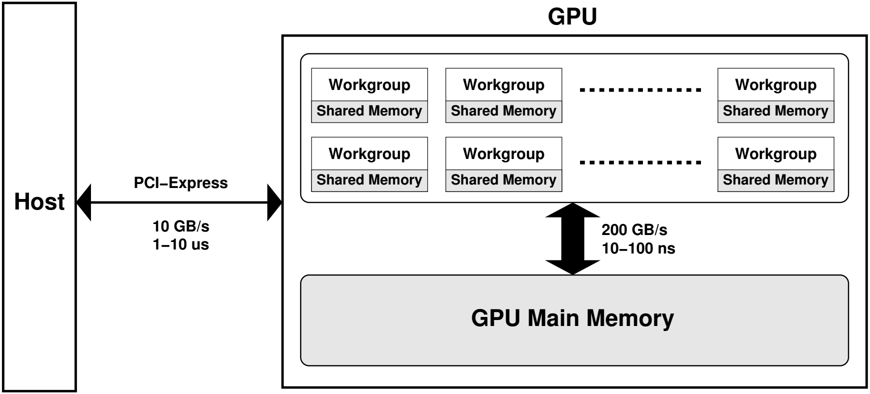
# Introduction

## Theoretical Peak Performance (FLOPs) per Byte of Memory Bandwidth



https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

# GPU Overview



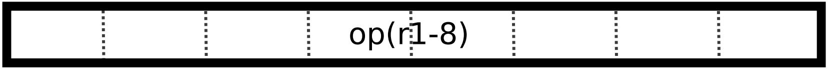
## Details

- Workgroups consist of 32-64 hardware threads
- Up to 24 hardware workgroups
- Shared memory small: approx. 32-64 KB

# GPU Overview

## Reminder: AVX

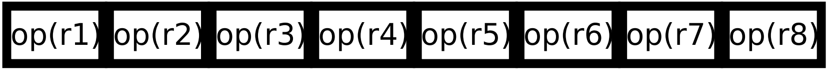
- One instruction for all elements of a vector register



op(r1-8)

## Single Instruction Multiple Threads (SIMT)

- One instruction for all threads in workgroup
- Each thread has separate registers
- Efficient if all threads execute the same instruction



op(r1) op(r2) op(r3) op(r4) op(r5) op(r6) op(r7) op(r8)

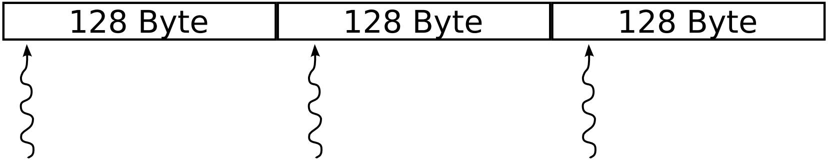
# GPU Overview

## GDDR5

- Optimized for throughput
- Channel width: multiple of 32 bits
- High bus width: 256 bits, 384 bits

## Structured Memory Access

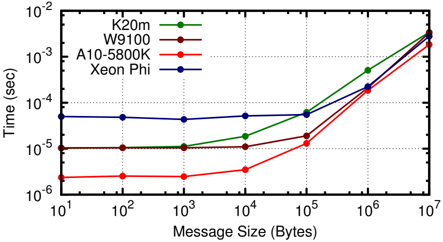
- Memory controllers use 32/64/128 byte transactions
- Partial transactions degrade effective bandwidth

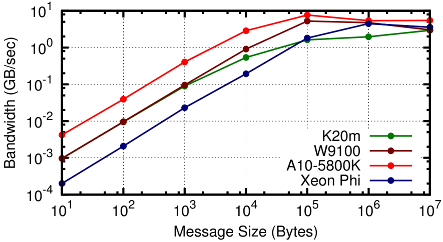


# GPU Overview

## Host-Device Communication

- PCI-Express v2:   8 GB/sec max
- PCI-Express v3: 16 GB/sec max
- PCI-Express v4: 32 GB/sec max
- Latency: about 10 $\mu$s


Time (sec)
10^-2
10^-3
10^-4
10^-5
10^-6
K20m
W9100
A10-5800K
Xeon Phi
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Message Size (Bytes)


Bandwidth (GB/sec)
10^1
10^0
10^-1
10^-2
10^-3
10^-4
K20m
W9100
A10-5800K
Xeon Phi
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Message Size (Bytes)