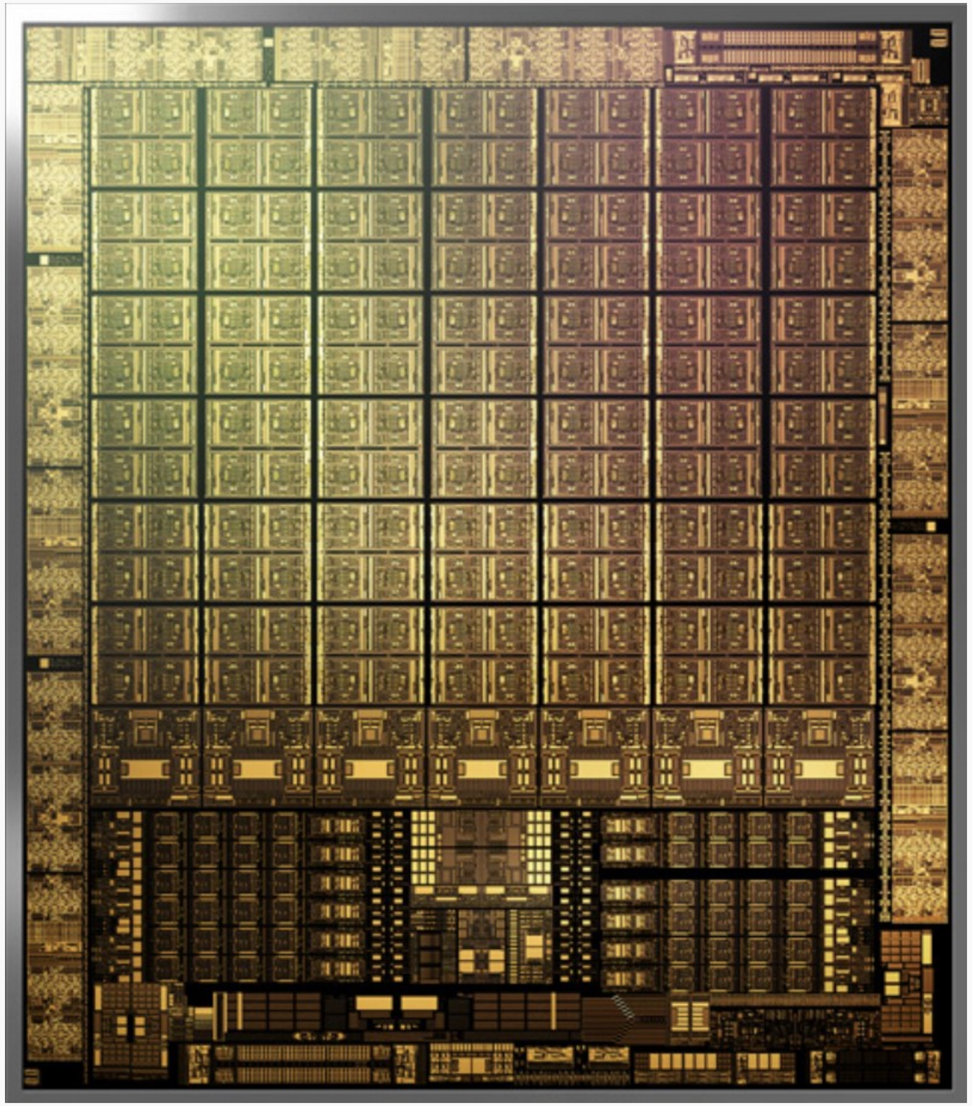
Arhitecturi Paralele
Introducere GPU

Lect. Dr. Ing. Cristian Chilipirea – cristian.chilipirea@mta.ro

NVIDIA Ampere

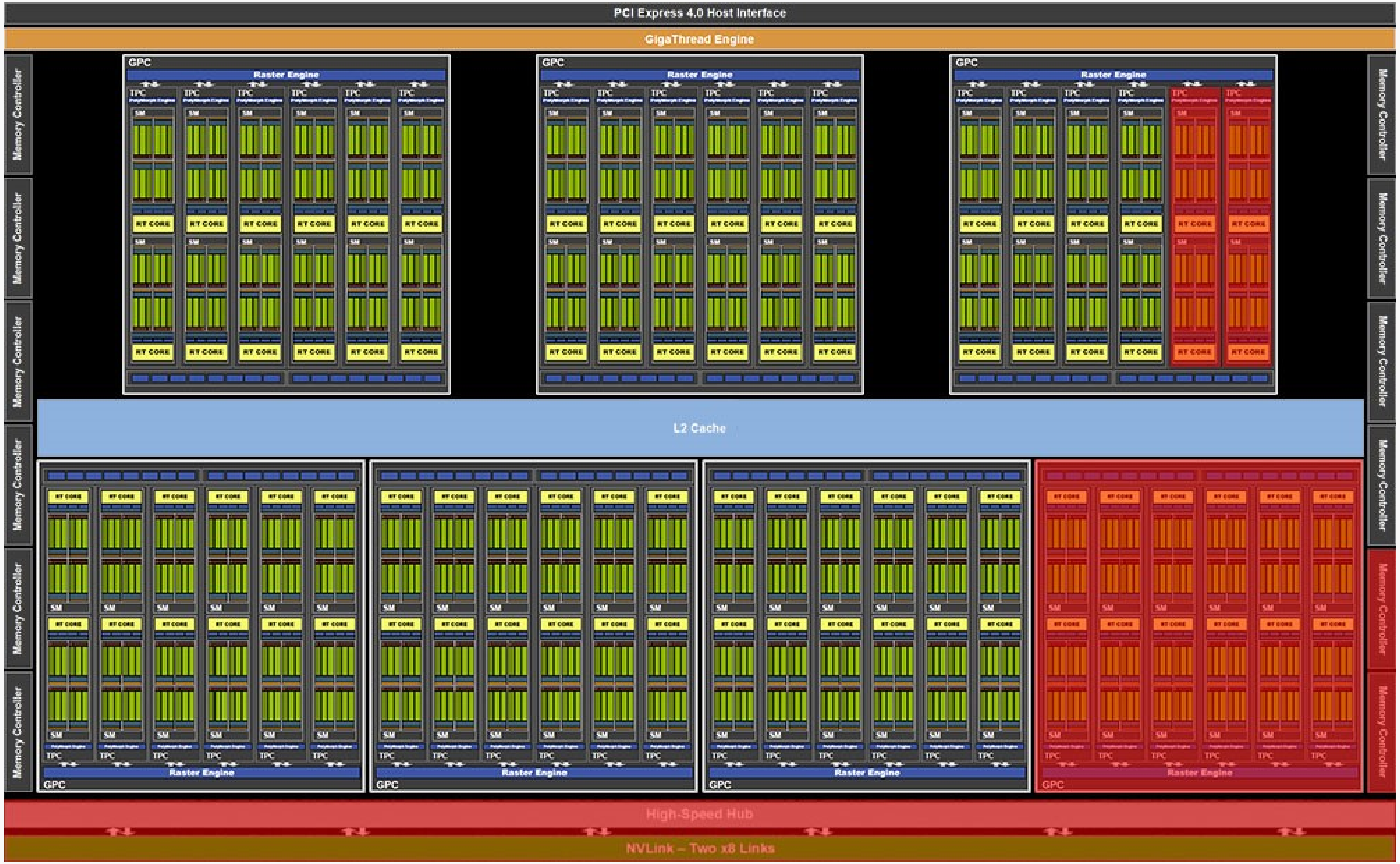


NVIDIA Ampere v2


PCI Express 4.0 Host Interface
GigaThread Engine
GPC
Raster Engine
TPC
SM
RT CORE
Memory Controller
L2 Cache
High-Speed Hub
NVLink – Four x4 Links

		GEFORCE RTX 3080 Ti	GEFORCE RTX 3080
Specificații GPU:	Nuclee NVIDIA CUDA®	10240	8704
	Frecvență Boost (GHz)	1.67	1.71
	Frecvență de bază (GHz)	1.37	1.44
Specificații memorie:	Configurație memorie standard	12 GB GDDR6X	10 GB GDDR6X
	Lățime interfață memorie	384 biți	320 biți
Tehnologii integrate:	Nuclee cu ray-tracing	Cea de-a doua generație	Cea de-a doua generație
	Nuclee Tensor	Cea de-a treia generație	Cea de-a treia generație
	Arhitectură NVIDIA	Ampere	Ampere

RTX 3080



PCI Express 4.0 Host Interface
GigaThread Engine
GPC
Raster Engine
TPC
SM
RT CORE
Memory Controller
L2 Cache
High-Speed Hub
NVLink – Two x8 Links

NVIDIA A40

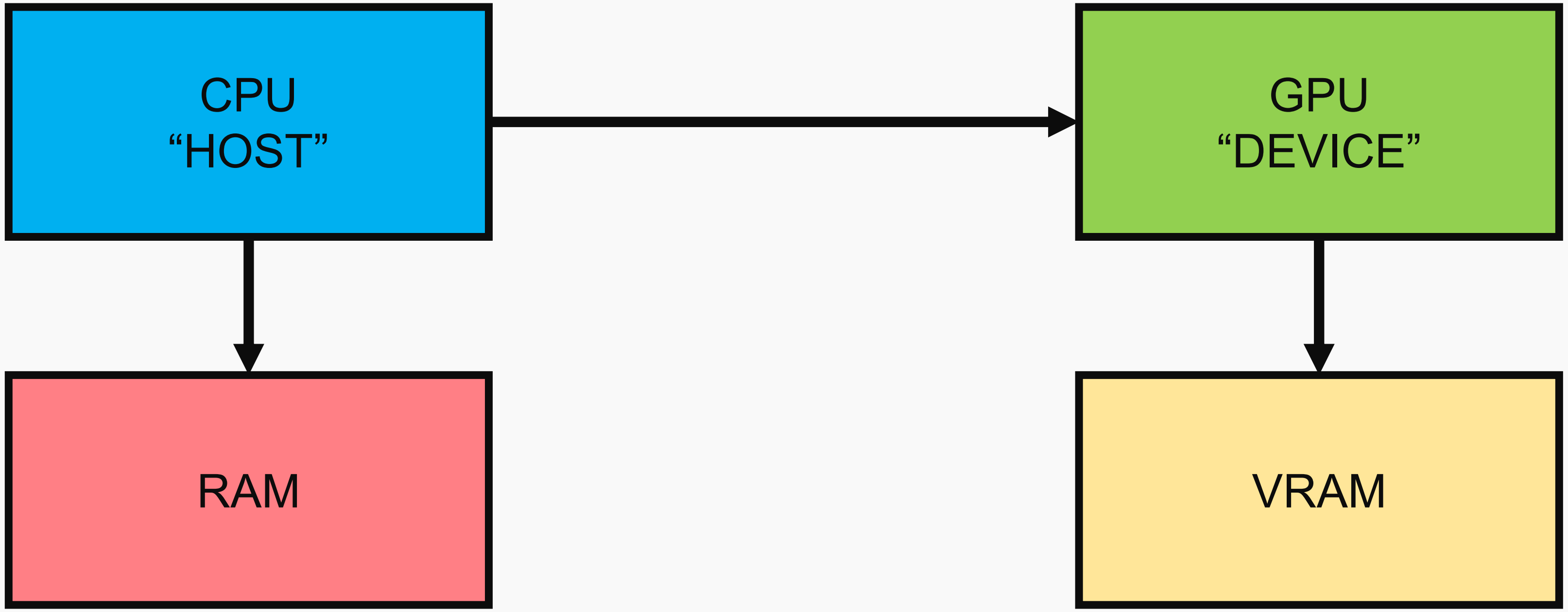
SPECIFICATIONS

GPU architecture	**NVIDIA Ampere architecture**
GPU memory	**48 GB GDDR6 with ECC**
Memory bandwidth	**696 GB/s**
Interconnect interface	**NVIDIA® NVLink® 112.5 GB/s (bidirectional)[3] PCIe Gen4 31.5 GB/s (bidirectional)**
NVIDIA Ampere architecture-based CUDA Cores	**10,752**
NVIDIA second-generation RT Cores	**84**
NVIDIA third-generation Tensor Cores	**336**

CPU vs GPU?

- Cores
- Frequency
- Core complexity

Arhitectura system heterogen

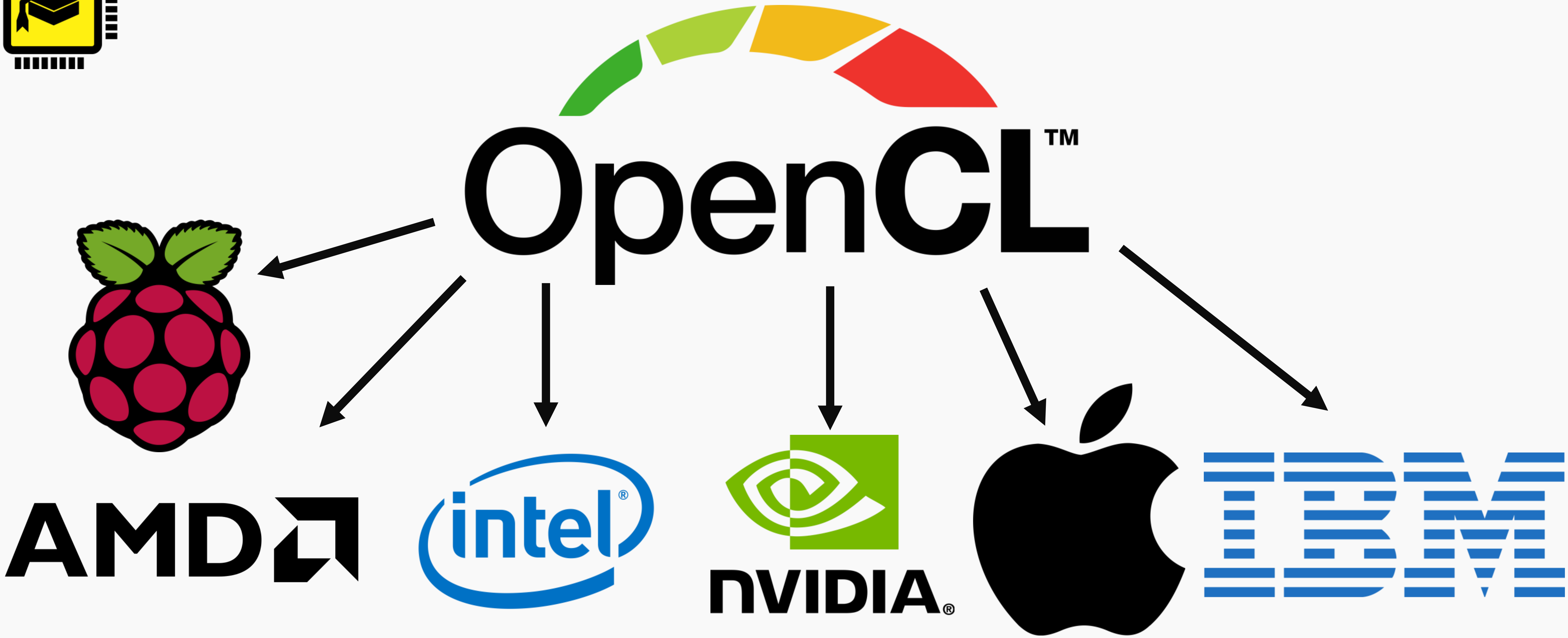


Typical Program

- CPU alocă memorie pe GPU (în VRAM)
- CPU copiază date din RAM în VRAM
- CPU pornește **kernelul** pe GPU
- CPU copiază date din VRAM pe RAM

Open Standard for Parallel Programming of Heterogeneous Systems

OpenCL™


OpenCL™
AMD
intel
NVIDIA
IBM

The industry's most pervasive, cross-vendor, open standard for low-level heterogeneous parallel programming

Desktop Creative Apps

Adobe, otoy, DASSAULT SYSTEMES, Modo, SONY, darktable, AUTODESK, CHAOSGROUP, CyberLink, ptc, Vegas Pro, RADEON PRORENDER, REALFLOW, GIMP, Capture One, ArcSoft, Blackmagicdesign, SILHOUETTE, blender, LuxCoreRender, SideFX, acdsee

Parallel Languages

SYCL, OpenACC (DIRECTIVES FOR ACCELERATORS), aparapi, PyOpenCL

Linear Algebra Libraries

SYCL-BLAS, ViennaCL, CLBlast

Machine Learning Libraries and Frameworks

Intel OPENVINO™, Intel clDNN, Xiaomi MACE (Mobile AI Compute Engine), SYCL-DNN, Caffe, NNAPI, Acuity (ML/AI Software Foundation) VeriSilicon, Qualcomm Neural Processing SDK for AI, Synopsis MetaWare EV, TI DL Library (TIDL), Arm Compute Library

Molecular Modelling Libraries

siremol.org, OpenMM, CHARMM (Chemistry at HARvard Macromolecular Mechanics), GROMACS (FAST. FLEXIBLE. FREE.), FOLDING@HOME, ForceBalance

Machine Learning Compilers

tvm.ai, GLOW, plaidML

Vision, Imaging and Video Libraries

FAST, OpenCV, OpenVX, VisionCpp, Metashape, Halide, FFmpeg

Math and Physics Libraries

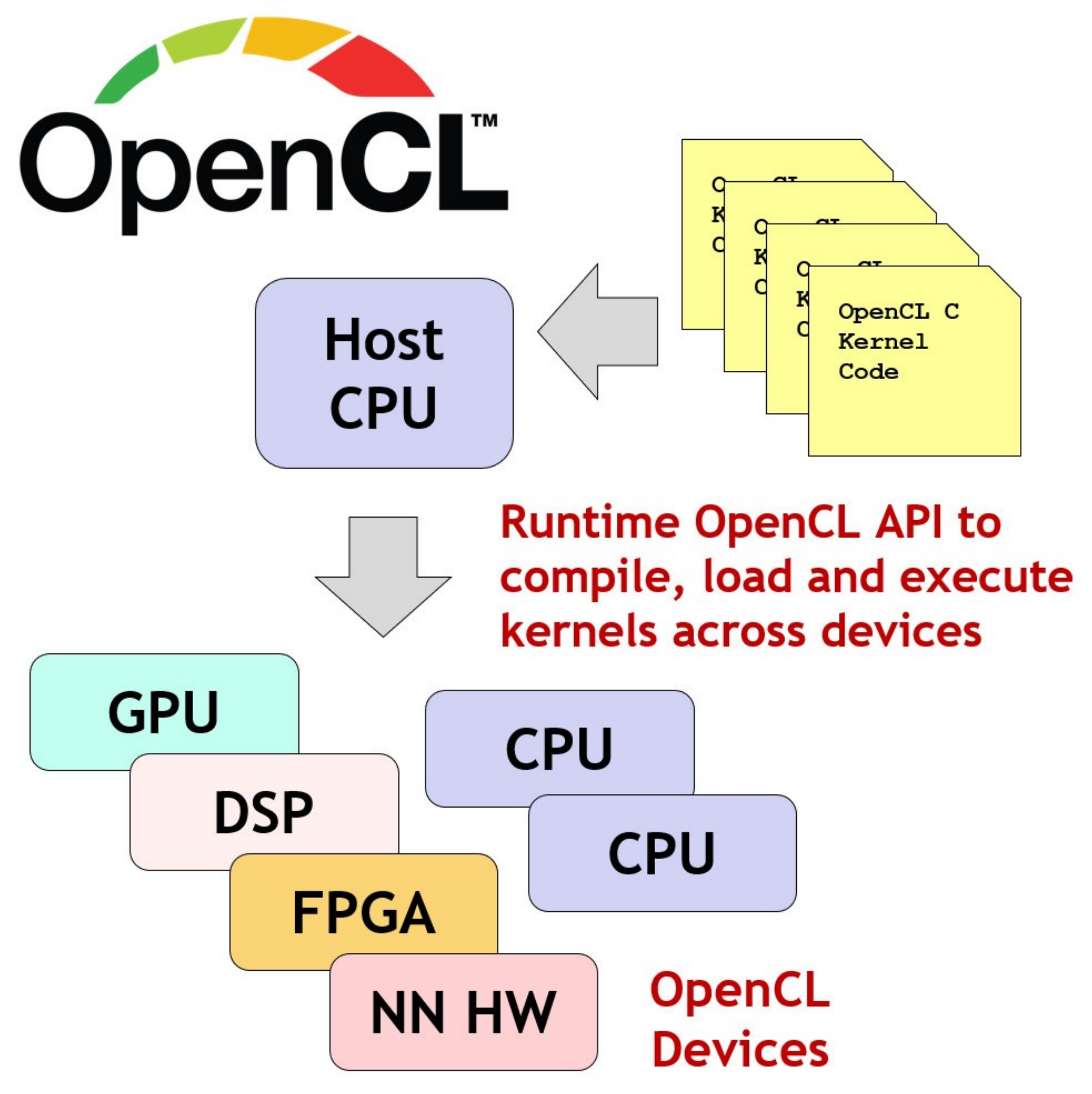
BULLET PHYSICS LIBRARY, GNU Octave, Wolfram Mathematica, ArrayFire (C, C++, Fortran), Matlab

---

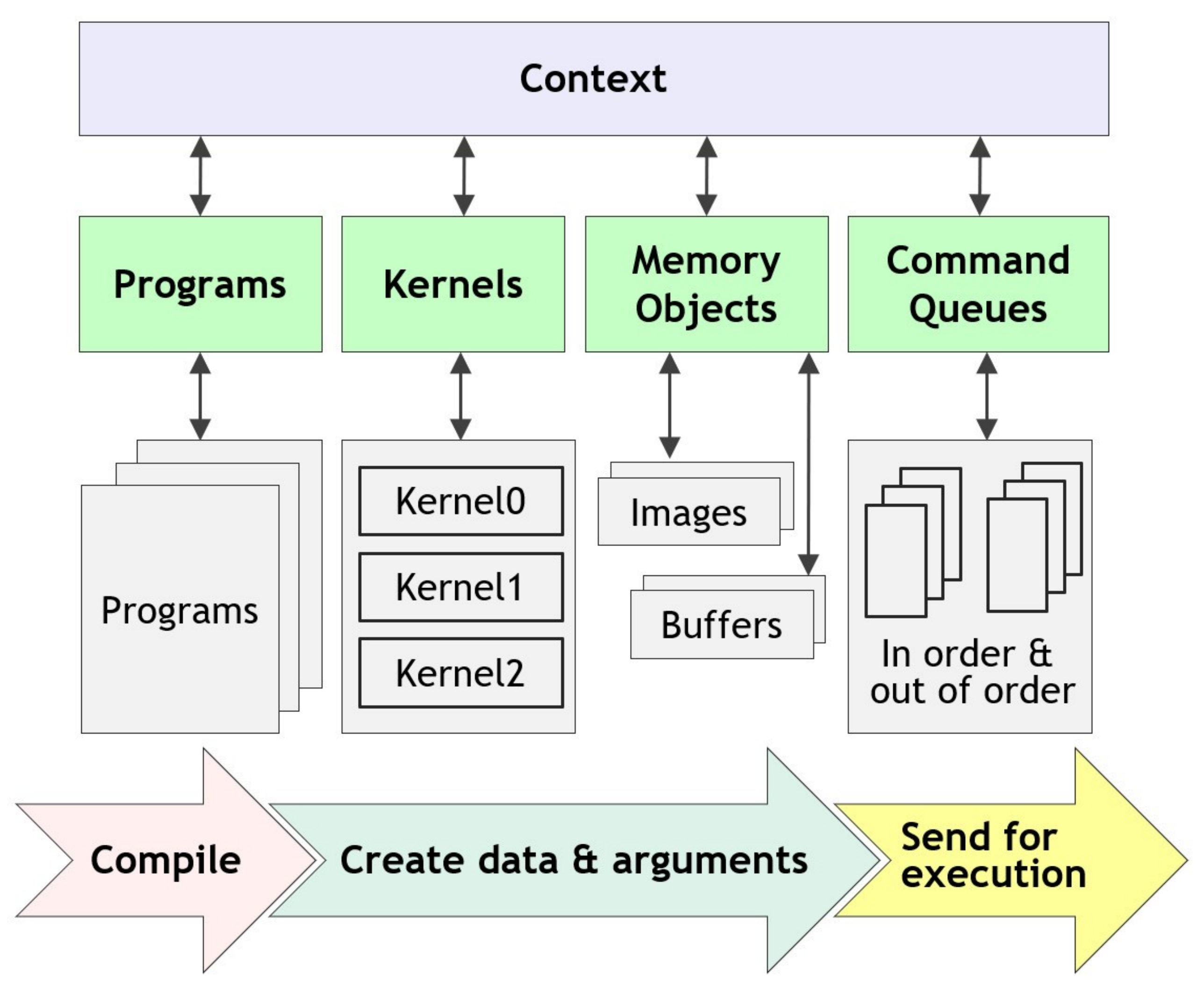
OpenCL™

AMD, arm, Imagination, intel, KALRAY, NVIDIA, MARVELL, MEDIATEK, QUALCOMM, SAMSUNG, ST, TEXAS INSTRUMENTS, VeriSilicon, XILINX

Accelerated Implementations


OpenCL™
OpenCL C
Kernel
Code
Host
CPU
Runtime OpenCL API to
compile, load and execute
kernels across devices
GPU
DSP
FPGA
NN HW
CPU
CPU
OpenCL
Devices

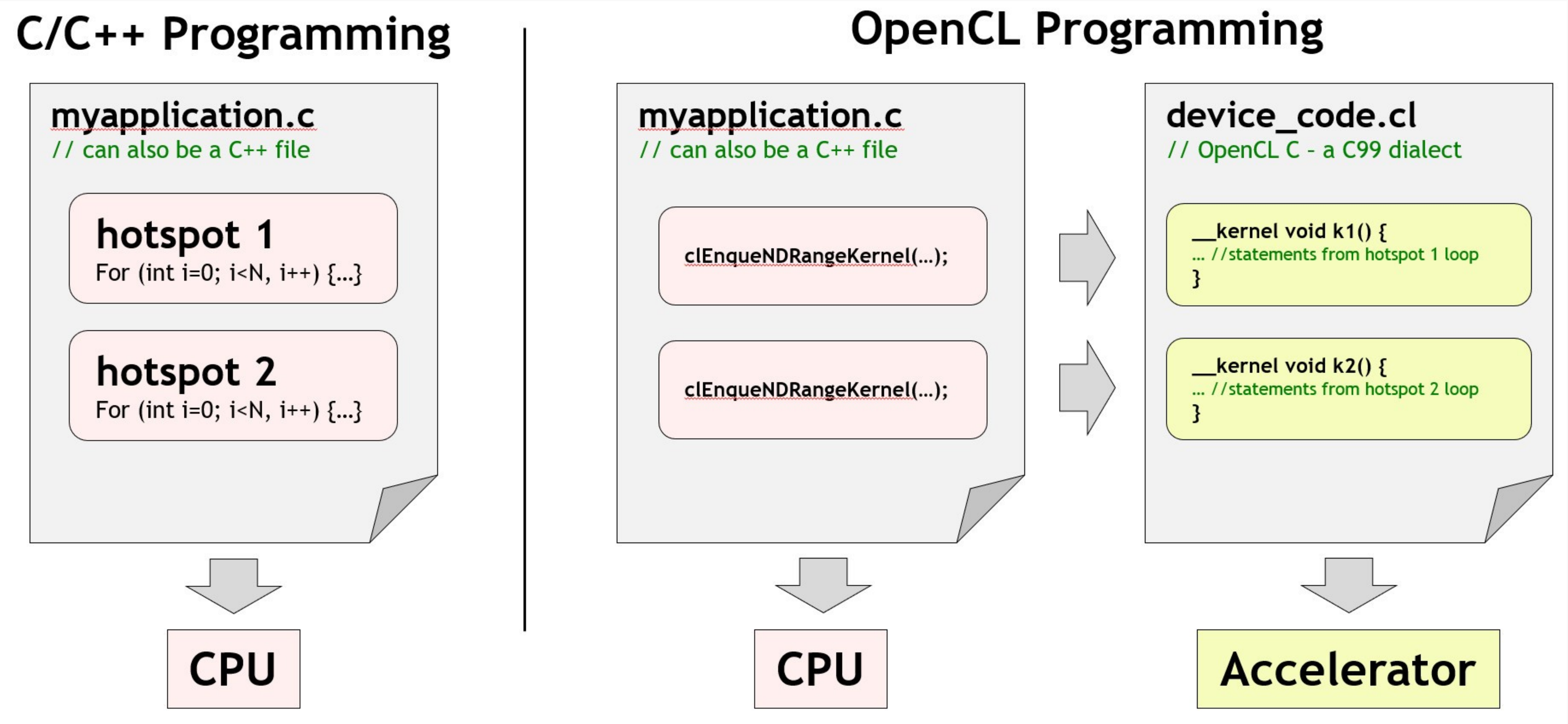
Sequence for Executing OpenCL Kernels



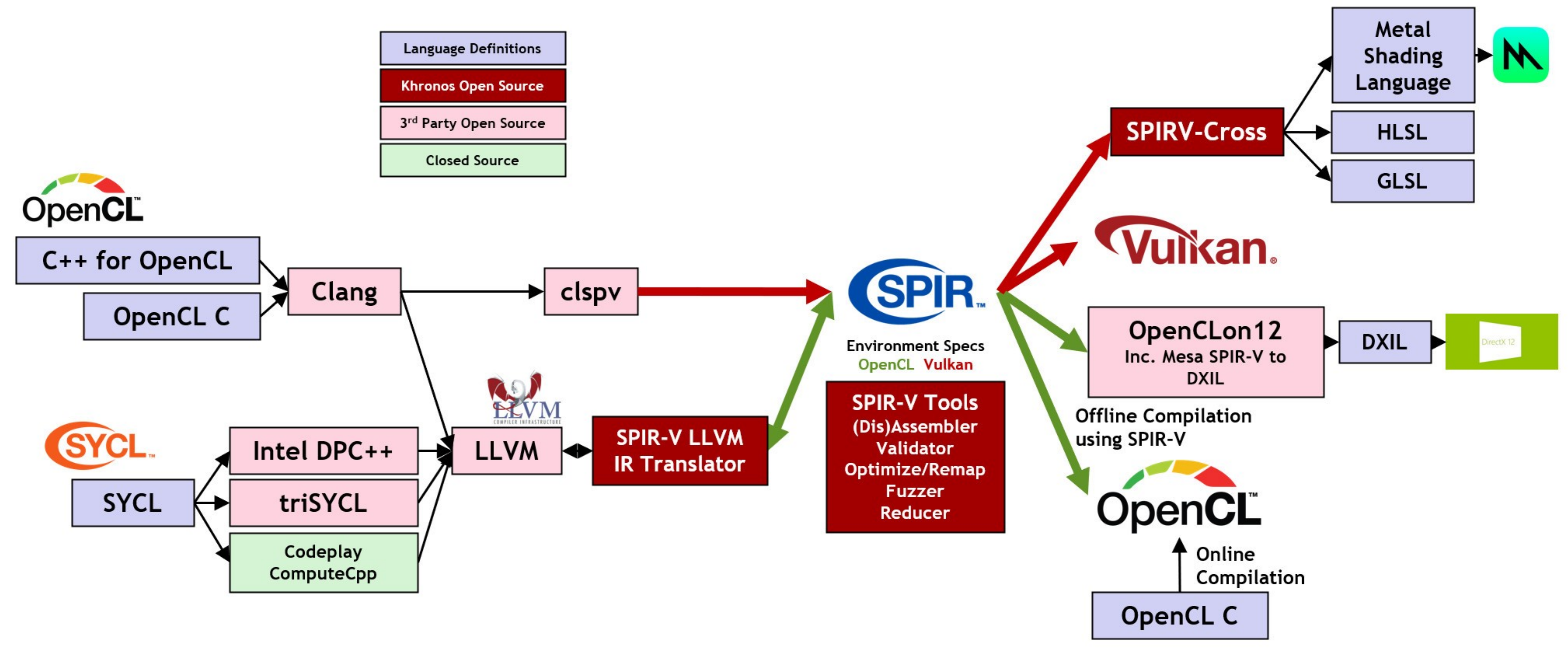
A complete sequence for executing an OpenCL program

- Query for available OpenCL platforms and devices
- Create a context for one or more OpenCL devices in a platform
- Create and build programs for OpenCL devices in the context
- Select kernels to execute from the programs
- Create memory objects for kernels to operate on
- Create command queues to execute commands on an OpenCL device
- *Enqueue* data transfer commands into the memory objects, if needed
- *Enqueue* kernels into the command queue for execution
- *Enqueue* commands to transfer data back to the host, if needed

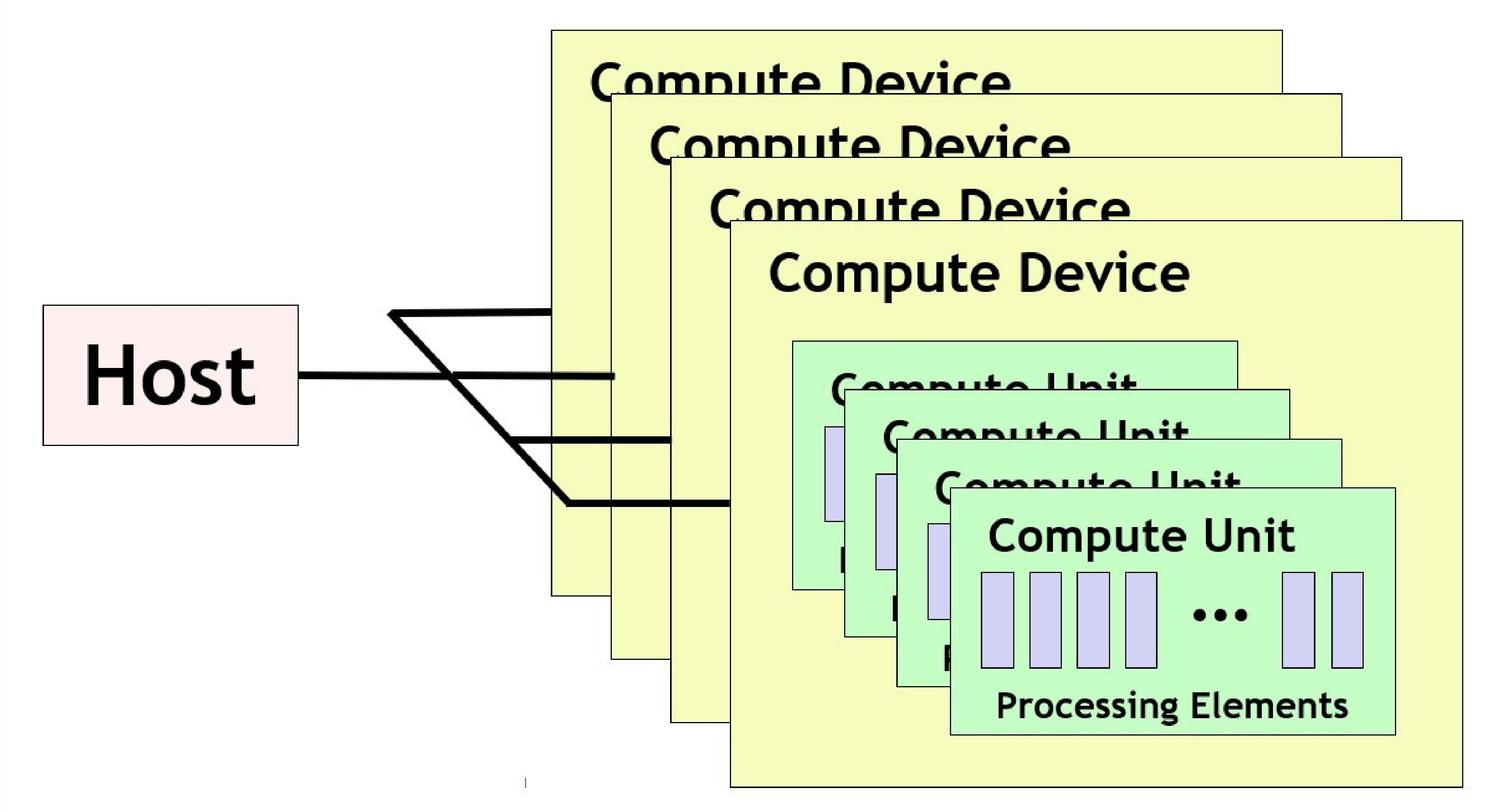
Traditional Versus OpenCL Programming Using OpenCL C Kernels



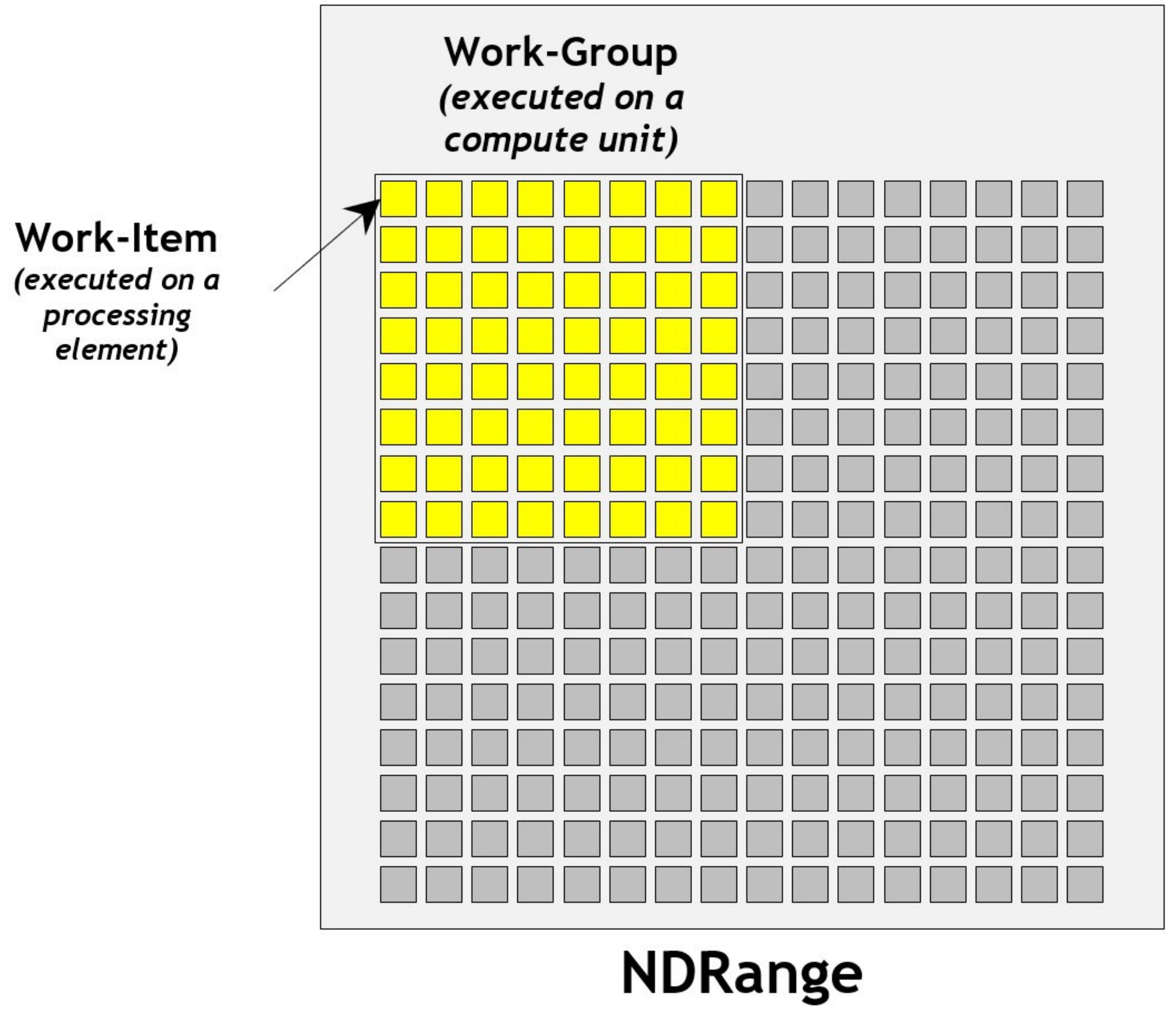
OpenCL Language Ecosystem Enabled With SPIR-V



OpenCL Platform Model



A 2D Image as an Example NDRange



OpenCL Memory Model