Гетерогенная архитектура HSA: экосистема для CPU/GPU/DSP

Илья Перминов

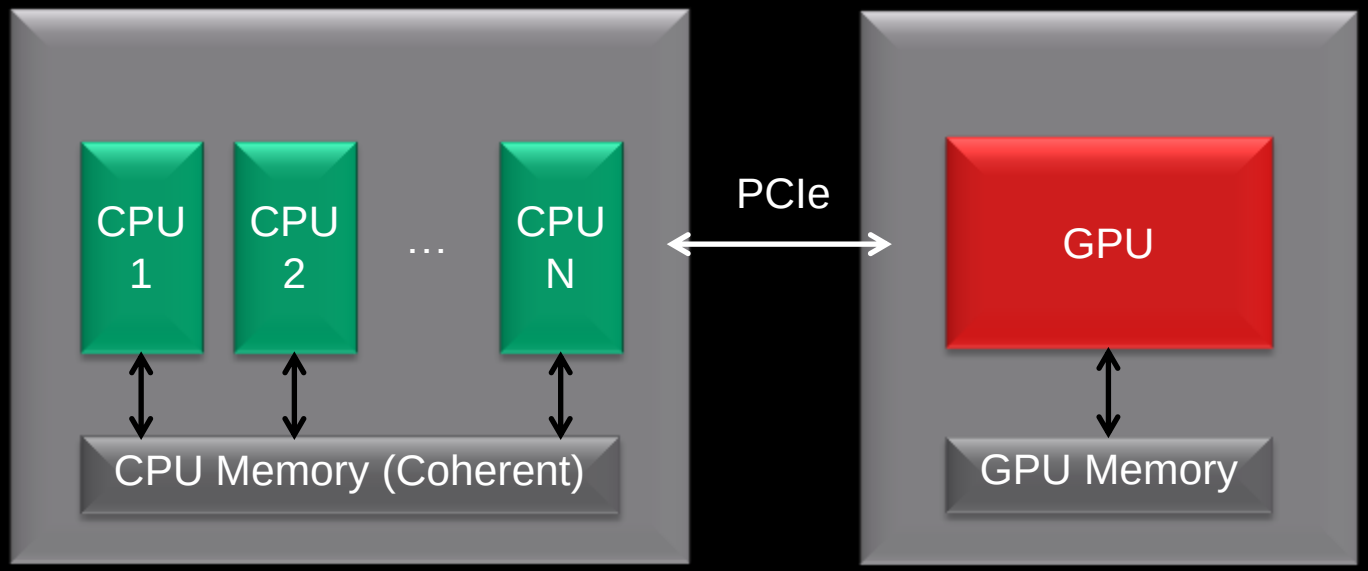
Тимур Палташев

23.10.2013

AMD

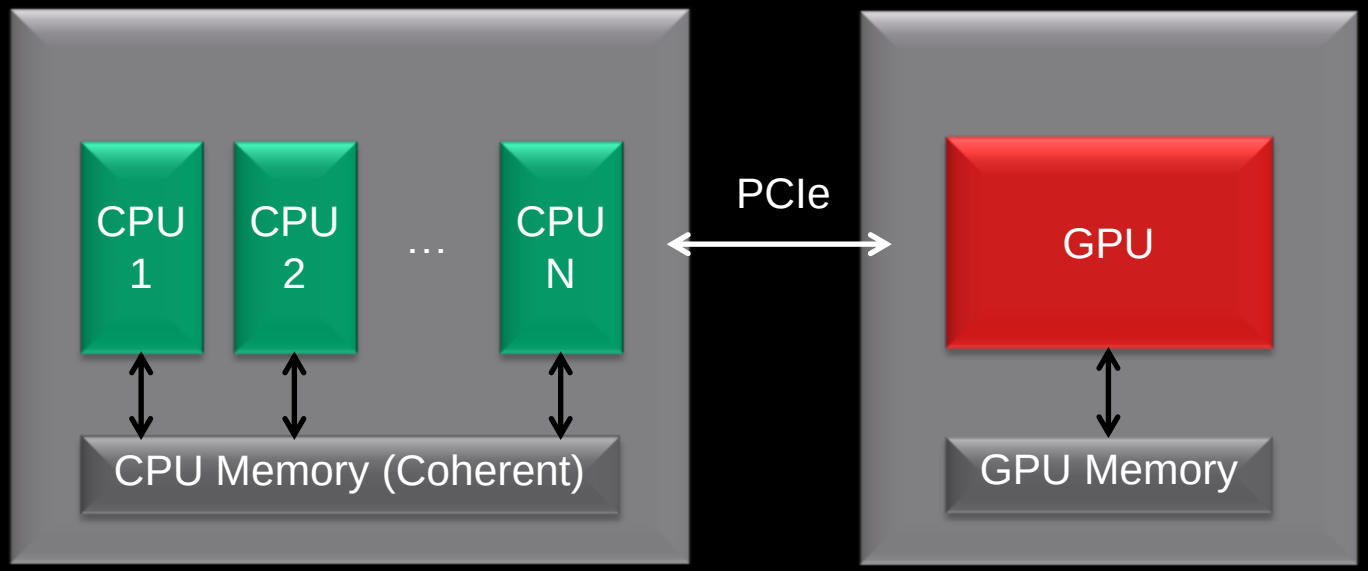
НИУ ИТМО

ЧТО ТАКОЕ HSA



HSA = Heterogeneous System Architecture

ЧТО ТАКОЕ HSA



Гетерогенные вычисления = GPGPU ?

ЧТО ТАКОЕ HSA

Гетерогенные вычисления = GPGPU ?

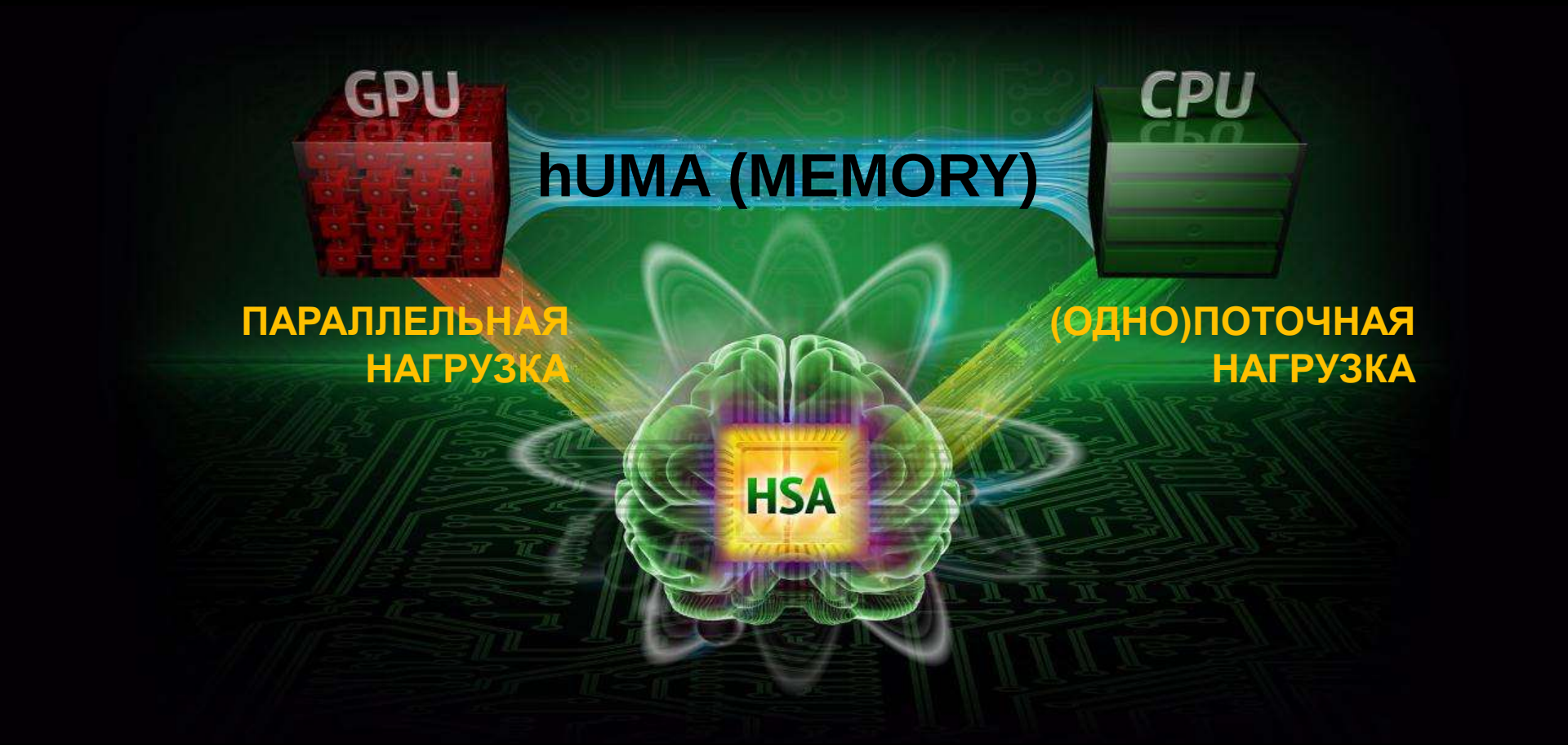
ЧТО ТАКОЕ HSA

Гетерогенные вычисления = GPGPU ?

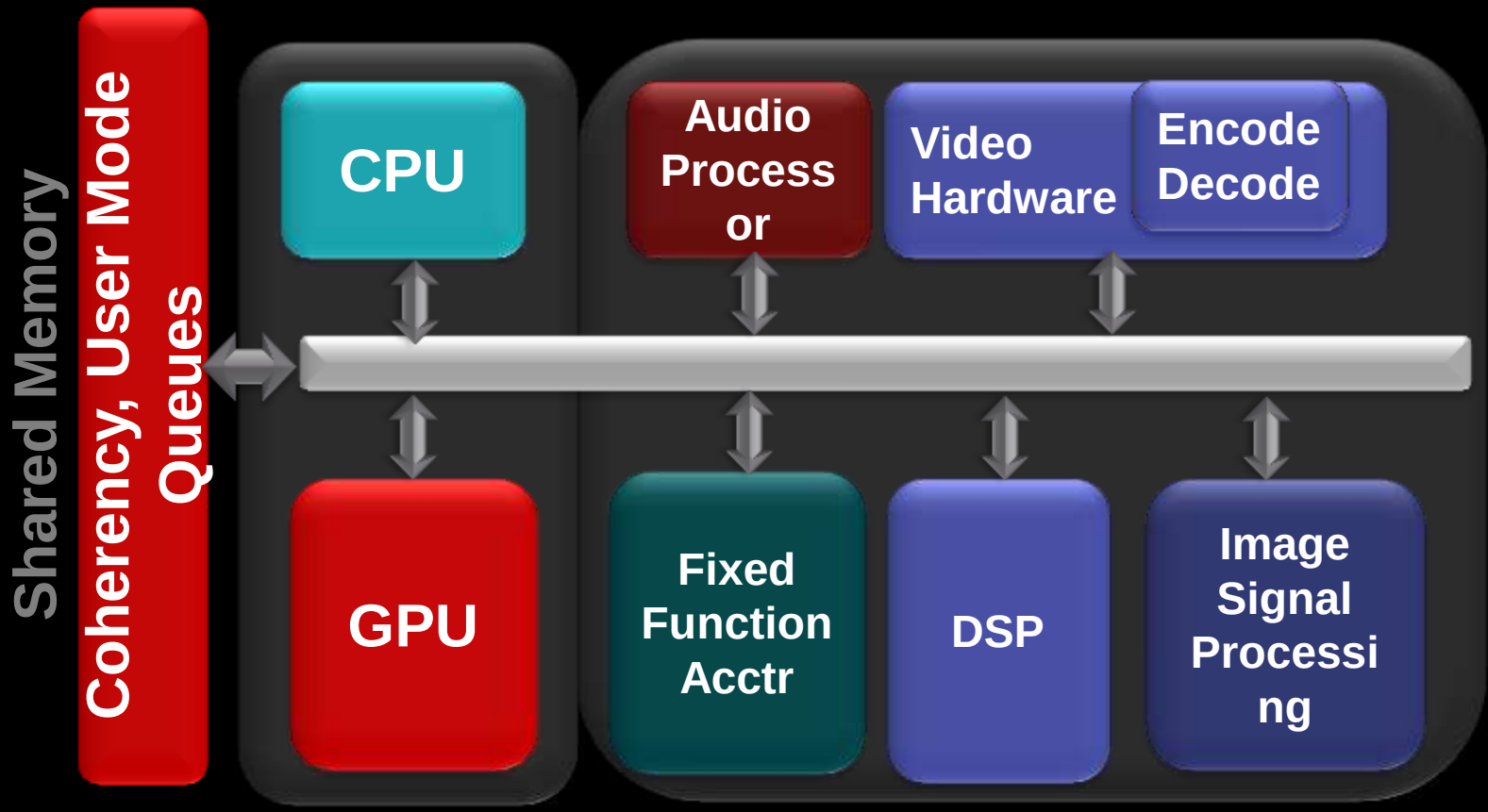
ЧТО ТАКОЕ HSA

Гетерогенные вычисления = GPGPU ?

HSA: HETEROGENEOUS SYSTEM ARCHITECTURE



HSA HIGH LEVEL ARCHITECTURE



COCTAB HSA FOUNDATION - 2013

Founders

AMD ARM imagination

MEDIATEK QUALCOMM SAMSUNG Texas Instruments

Promoters

LG Electronics

Supporters

Sandia National Laboratories Arteris codeplay FABRICENGINE MULTICORE WARE

Contributors

apical CEVA The DSP Powerhouse omp SONICS MARVELL SONY make.believe

ANALOG DEVICES ST life.augmented Swarm64 <symbio> tensilica VIVANTE ST ERICSSON

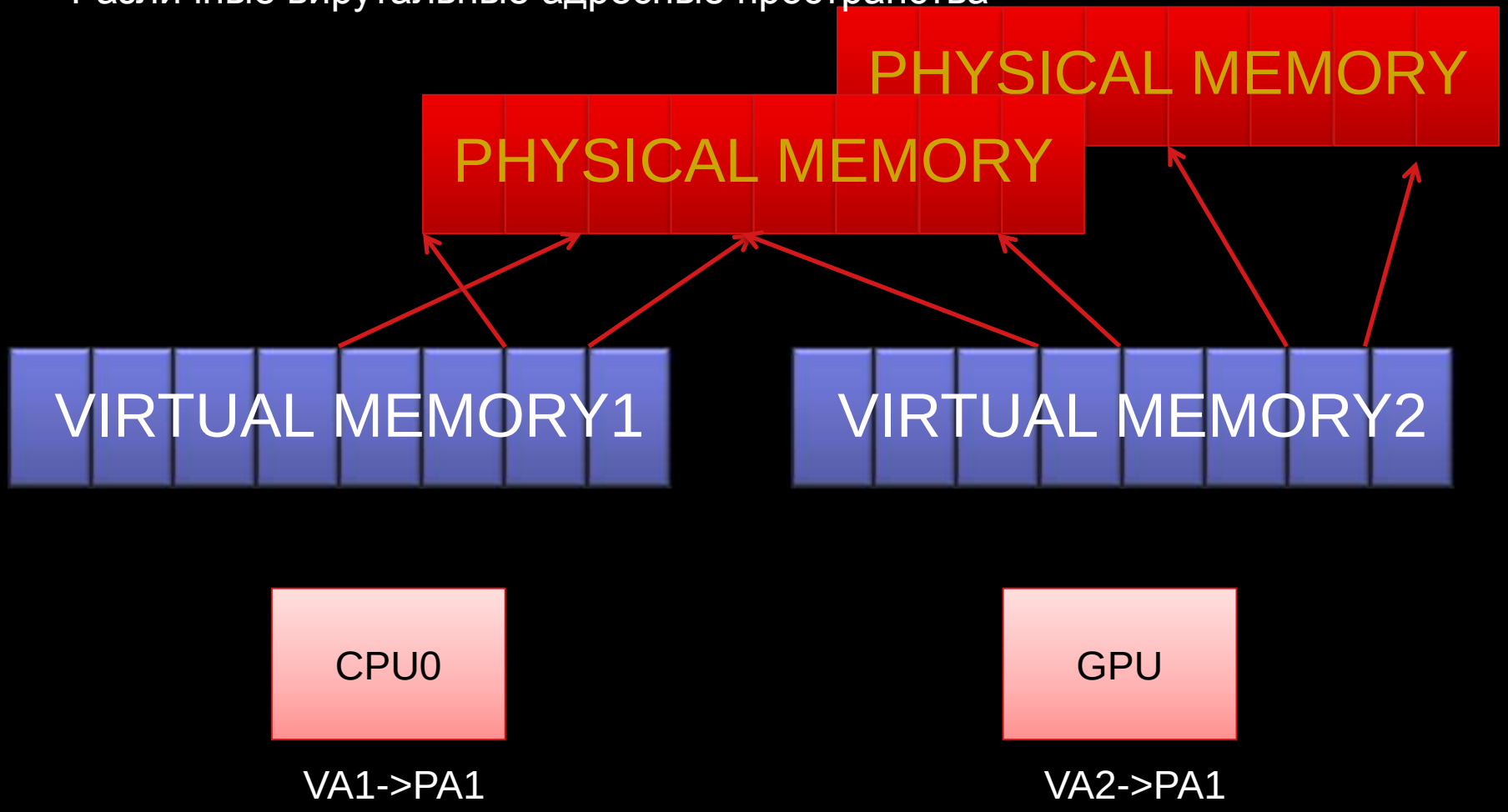
SYNOPSYS

Academic

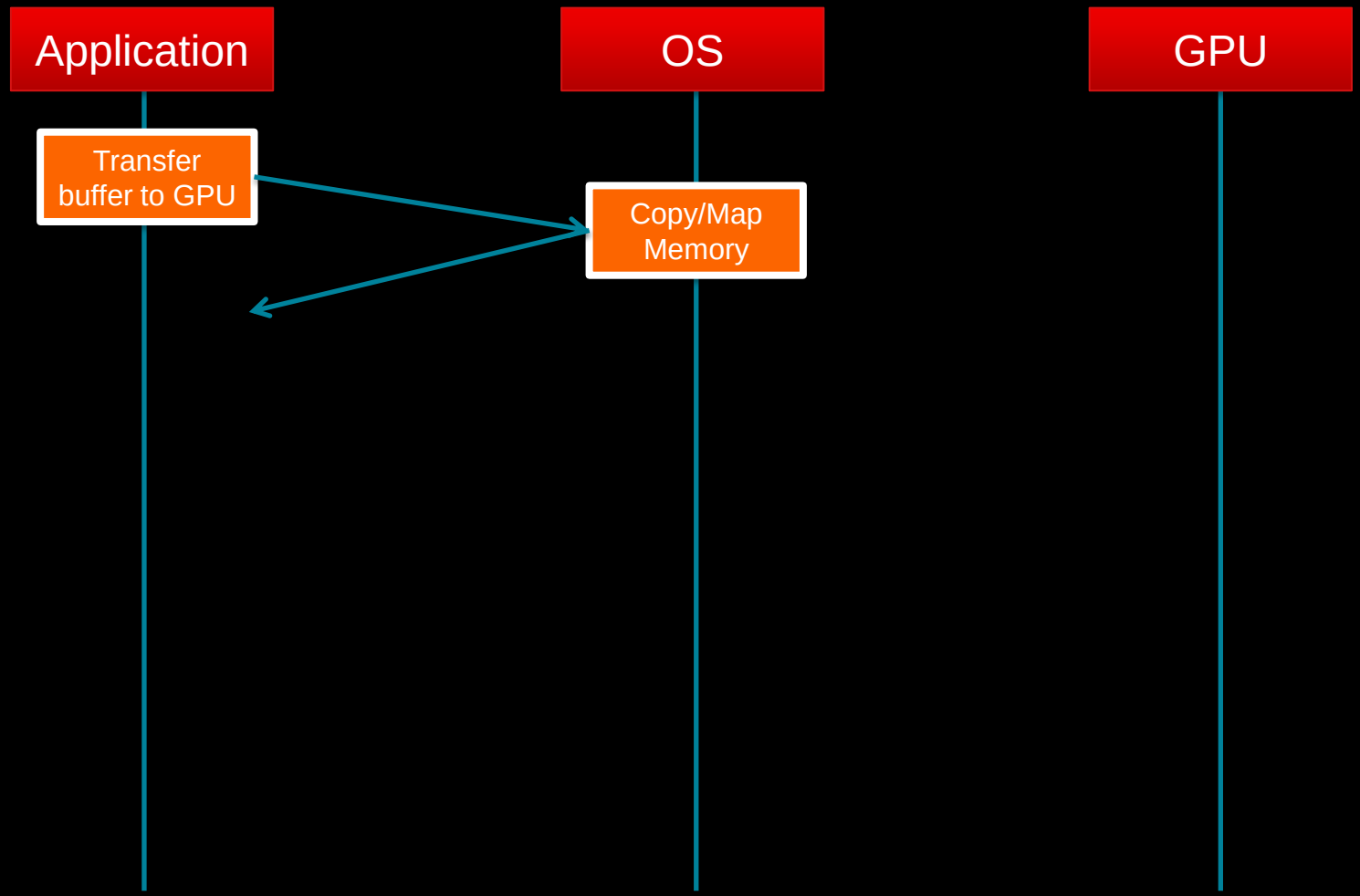
NTHU Programming Language Lab NTHU System Software Lab University of BRISTOL THE UNIVERSITY of EDINBURGH informatics ILLINOIS COMPUTER SCIENCE

SHARED VIRTUAL MEMORY (TODAY)

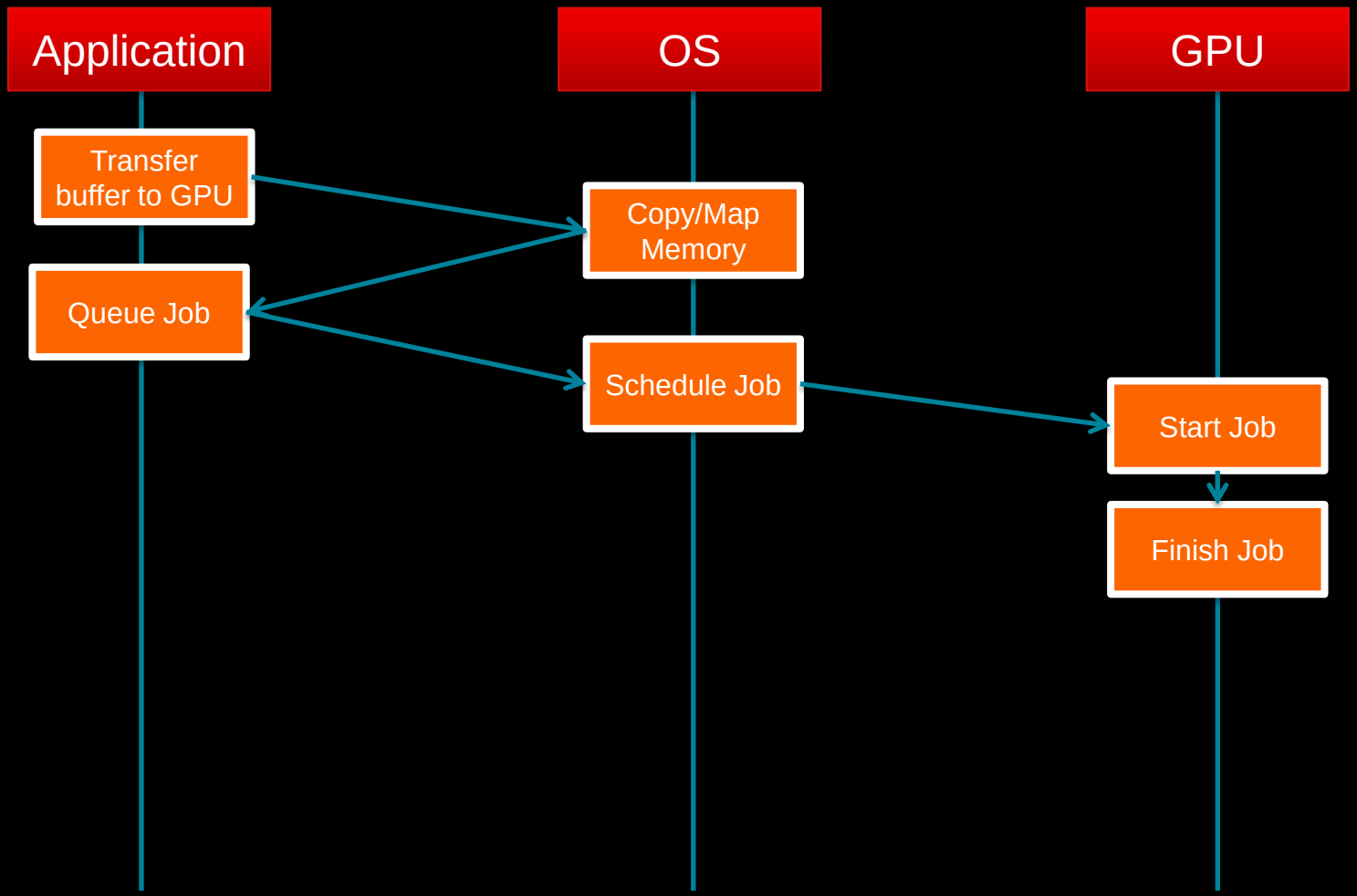
- Различные вирутальные адресные пространства



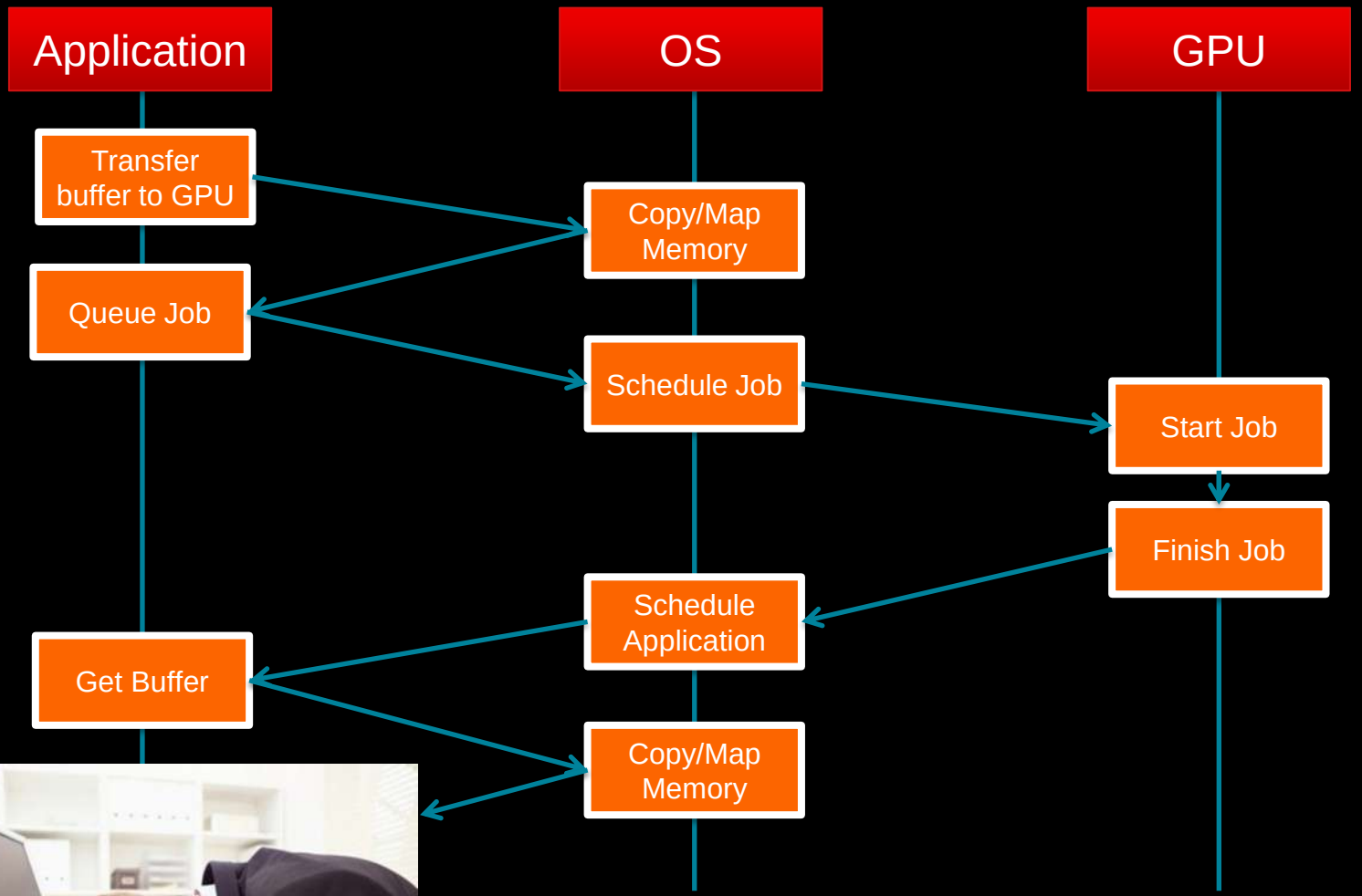
ЧТО МЫ ИМЕЕМ СЕГОДНЯ

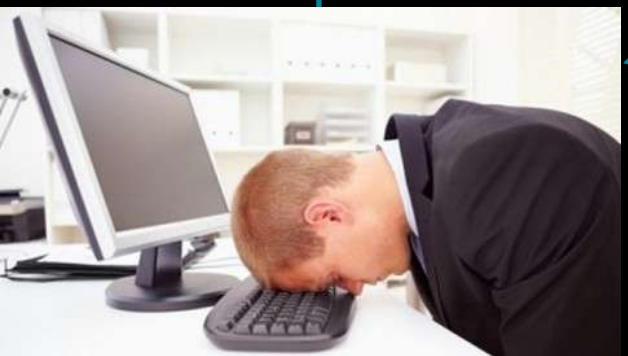


ЧТО МЫ ИМЕЕМ СЕГОДНЯ

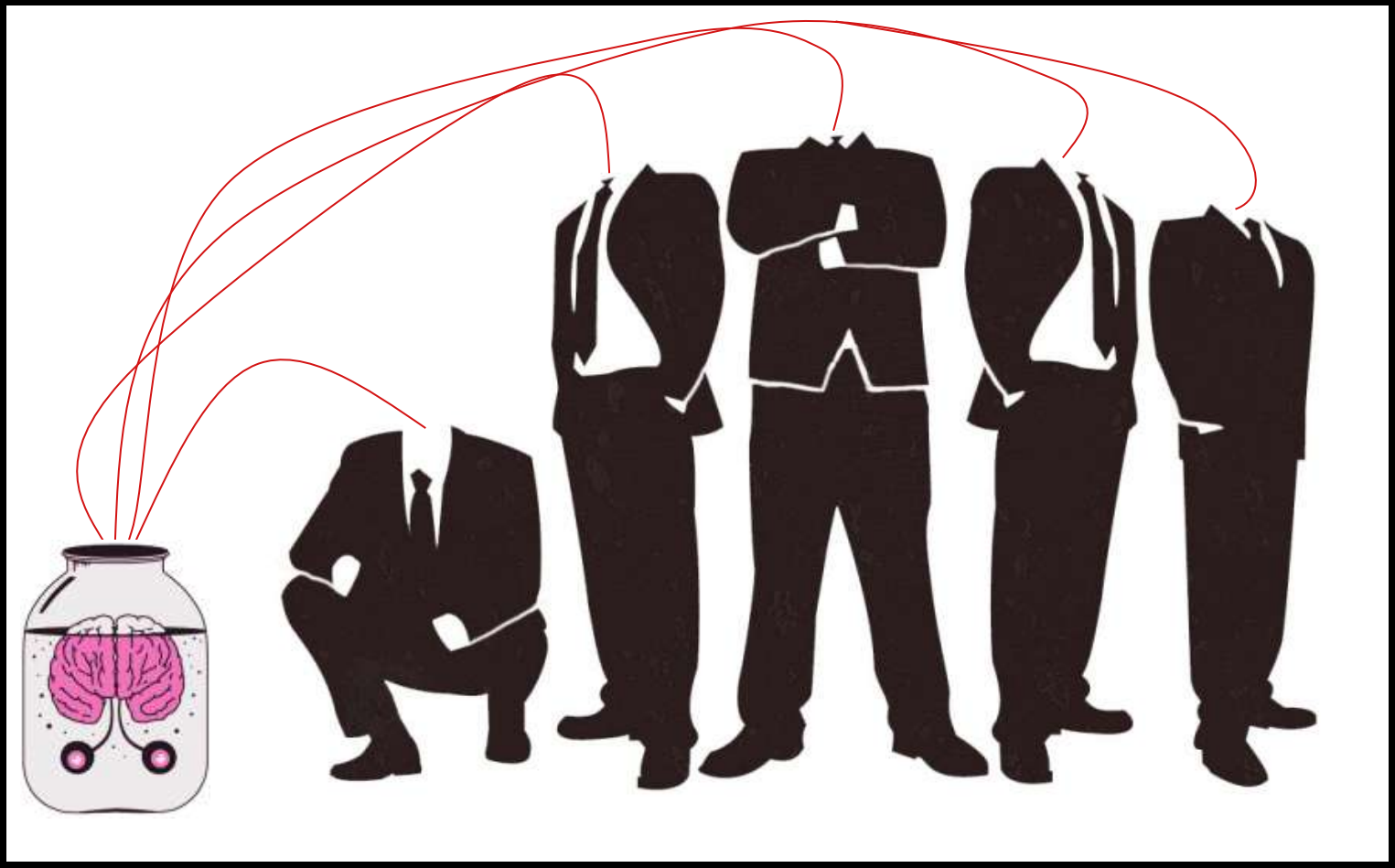


ЧТО МЫ ИМЕЕМ СЕГОДНЯ





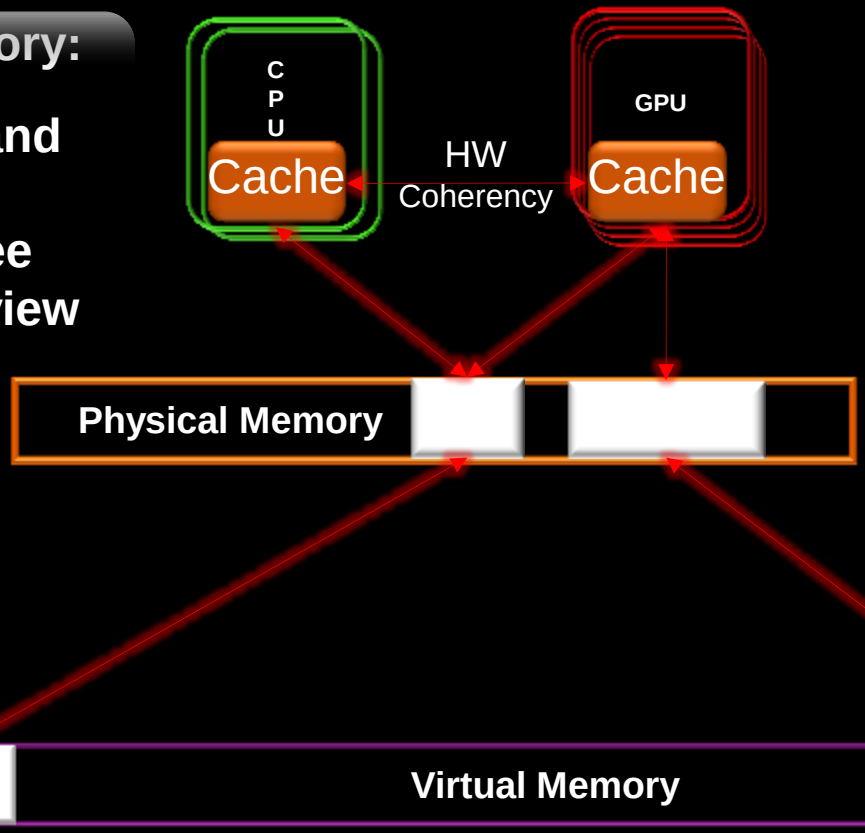
ОБЩАЯ КОГЕРЕНТНАЯ ПАМЯТЬ (HUMA)



HSA KEY FEATURES

Coherent Memory:

Ensures CPU and GPU caches both see an up-to-date view of data



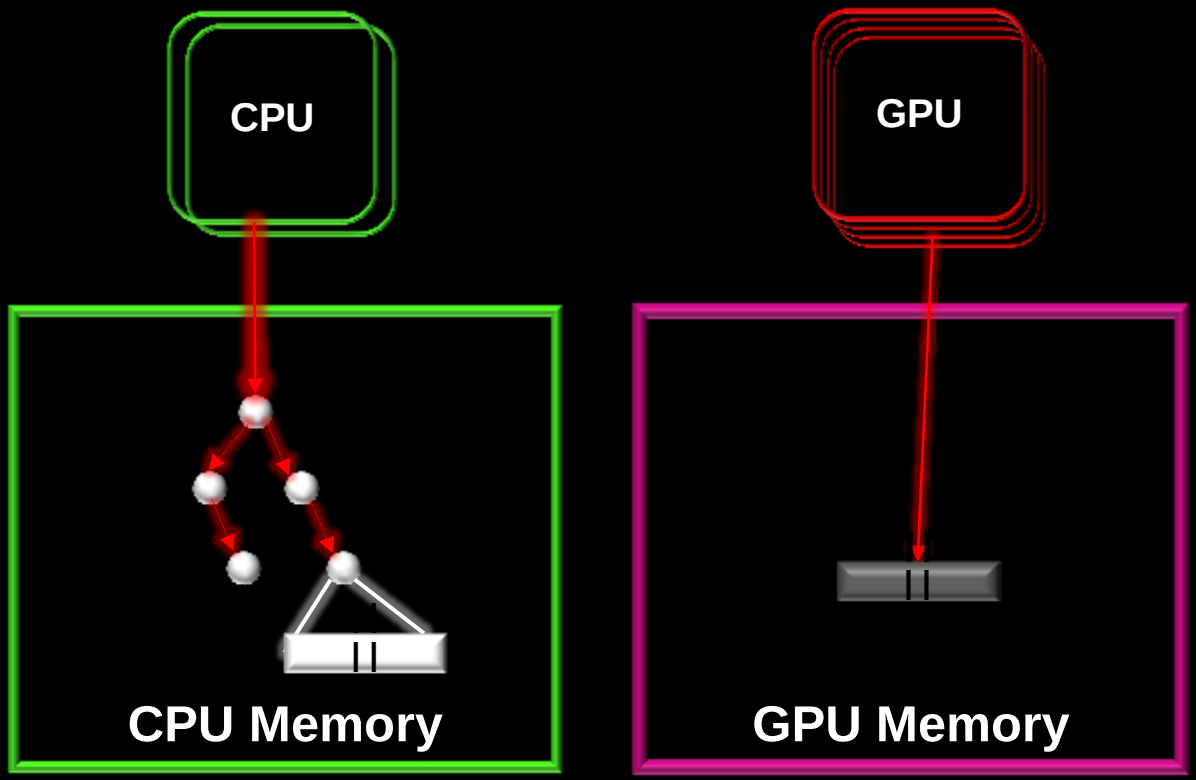
Pageable memory:

The GPU can seamlessly access virtual memory addresses that are not (yet) present in physical memory

Entire memory space:
Both CPU and GPU can access and allocate any location in the system's virtual memory space

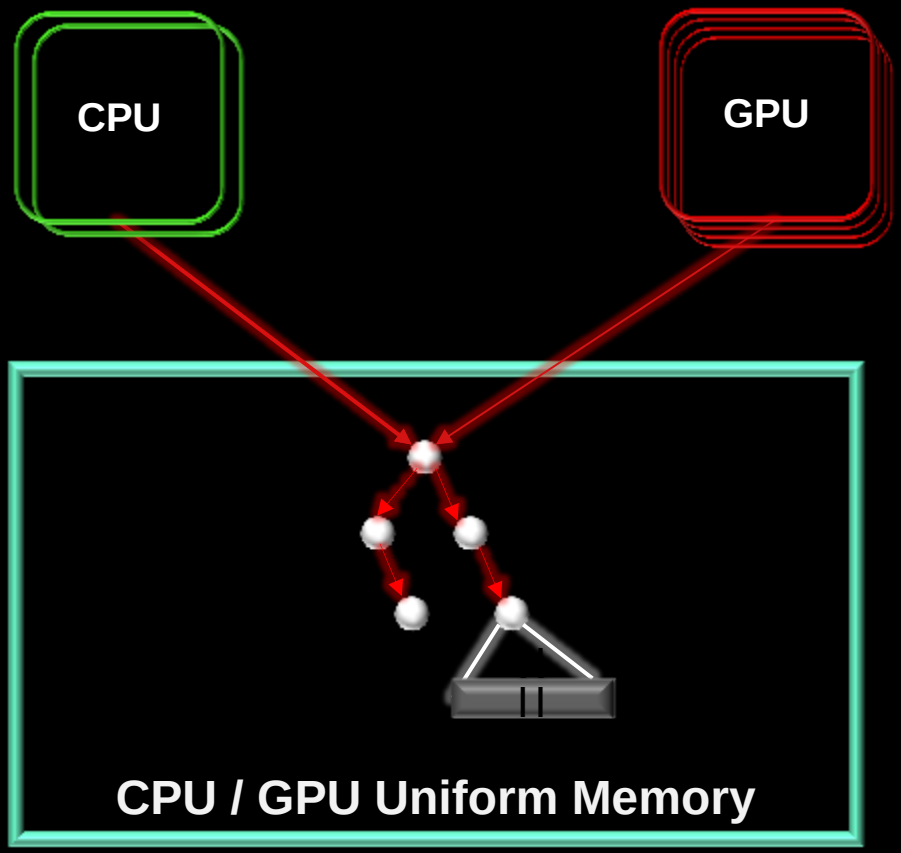
Без HSA

- CPU explicitly copies data to GPU memory
- GPU completes computation
- CPU explicitly copies result back to CPU memory

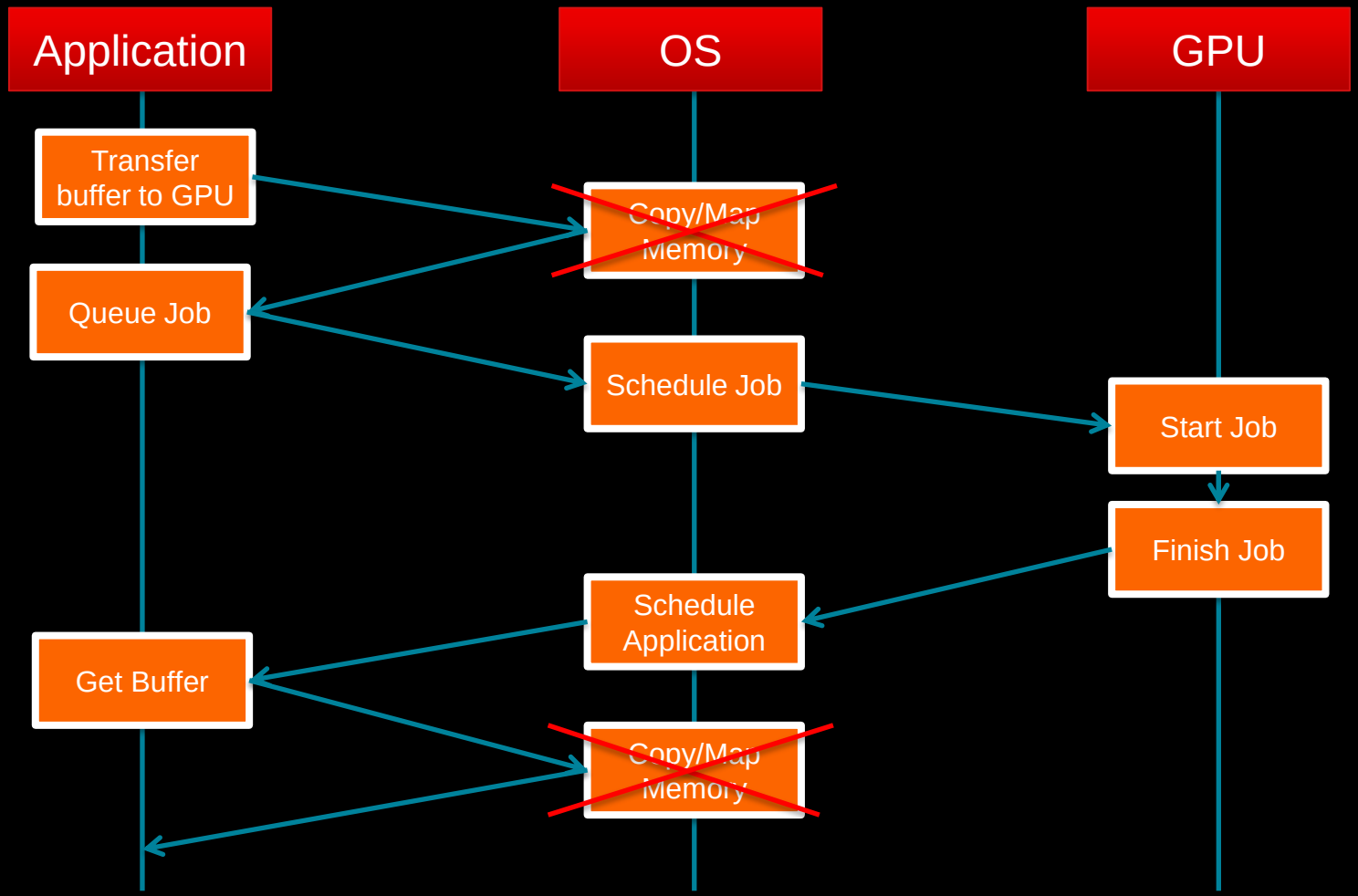


HSA

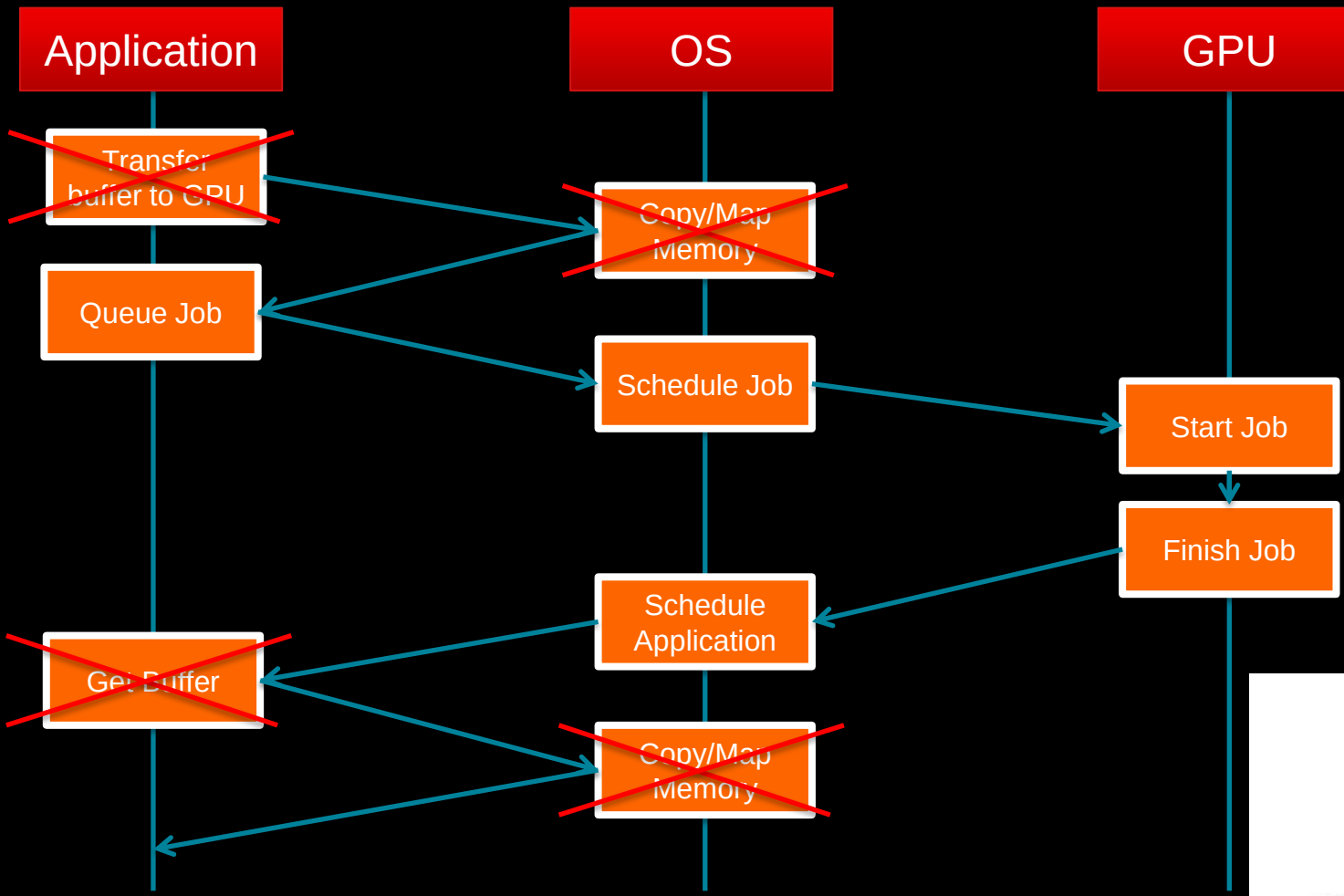
- CPU simply passes a pointer to GPU
- GPU complete computation
- CPU can read the result directly – no copying needed!



ОБЩАЯ ПАМЯТЬ

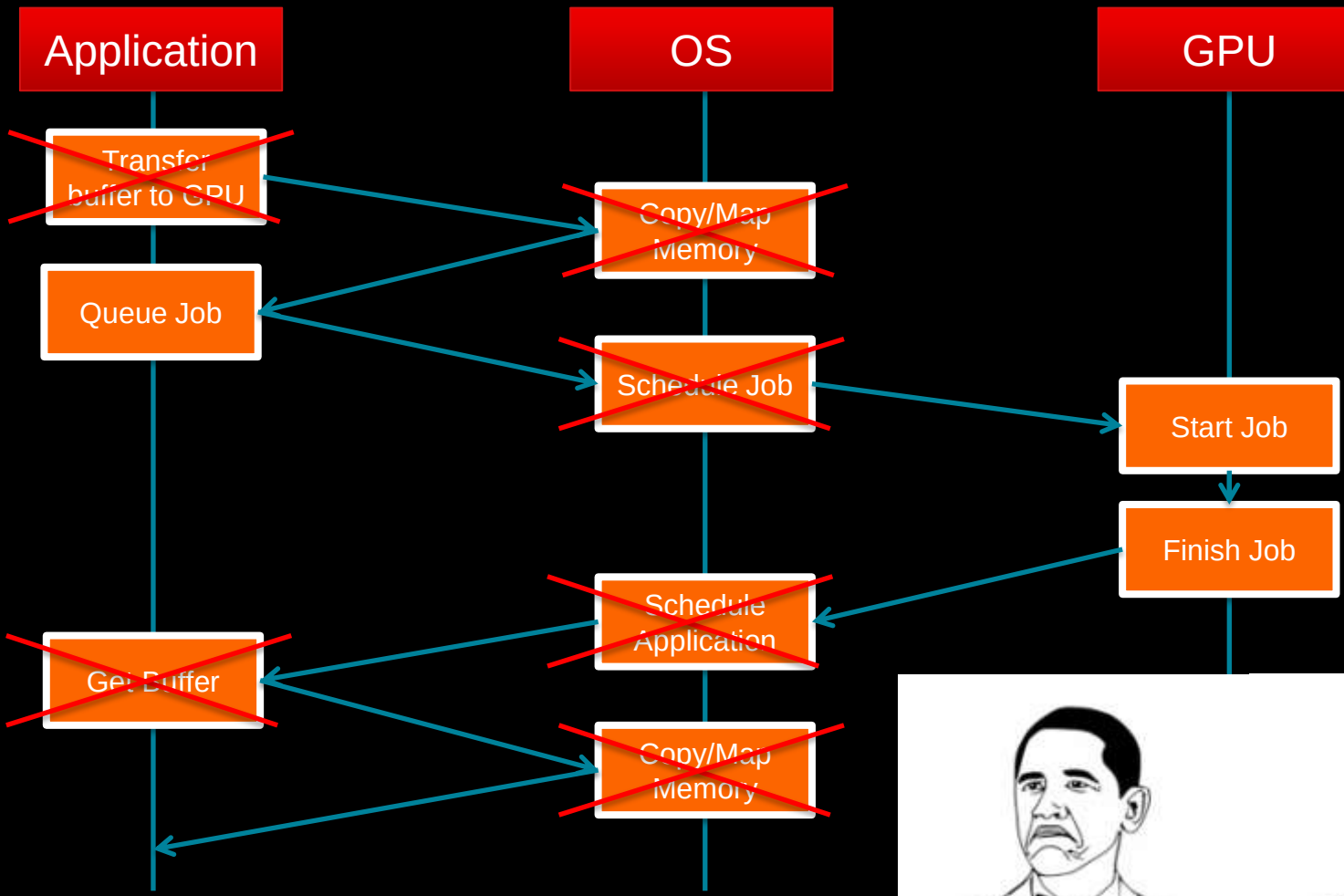


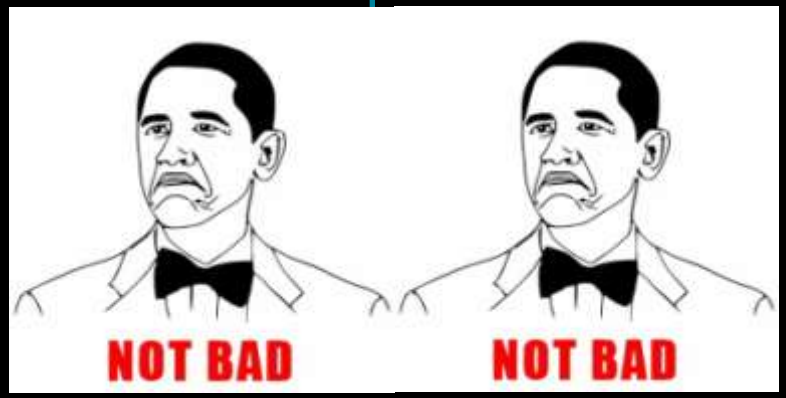
ОБЩАЯ КОГЕРЕНТНАЯ ПАМЯТЬ



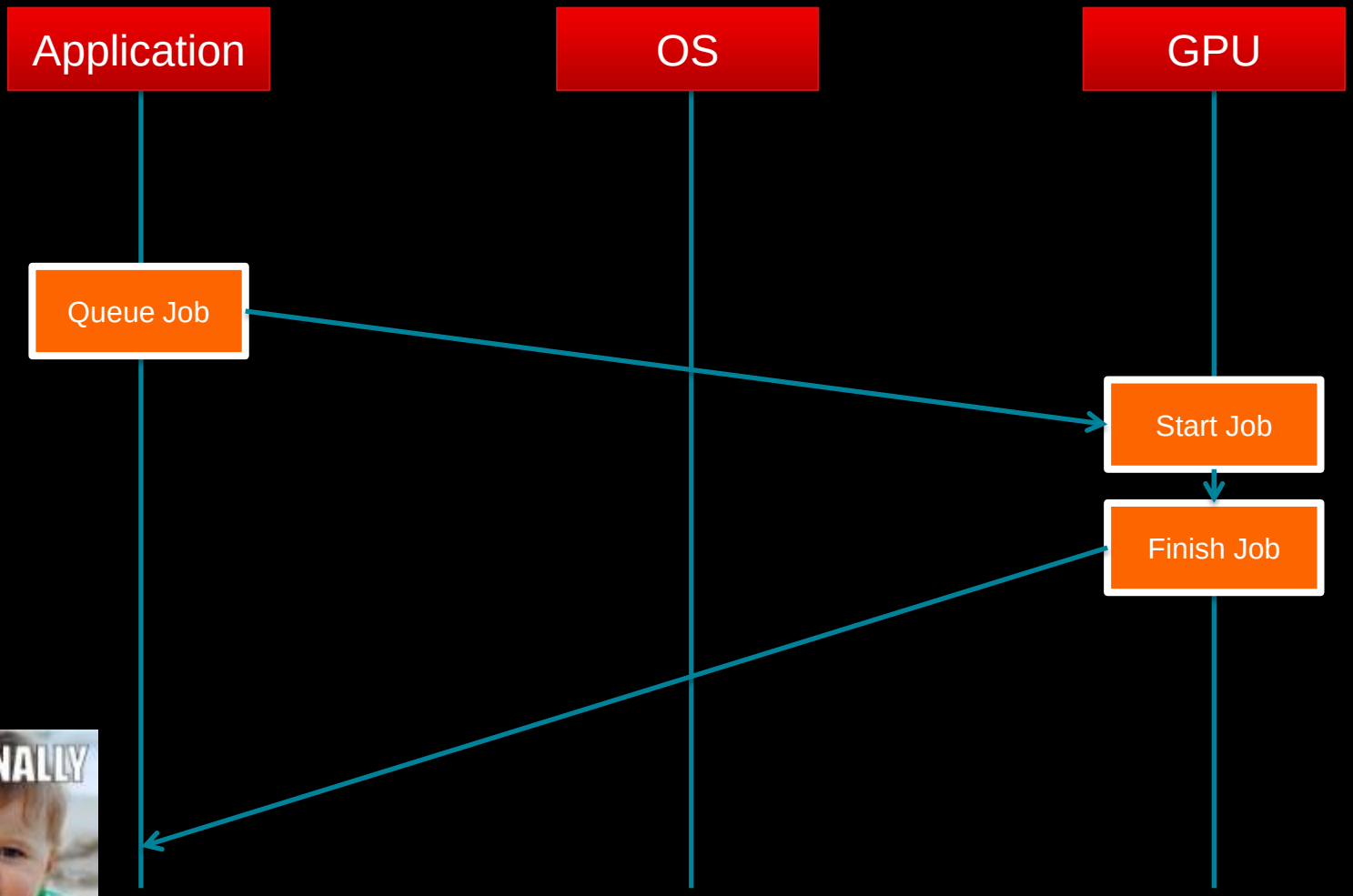
УПРАВЛЕНИЕ ОЧЕРЕДЯМИ ЗАДАЧ

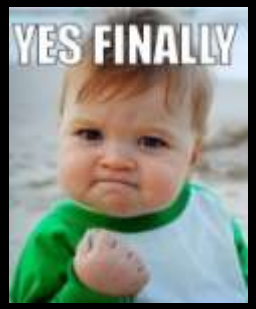
ДИСПЕТЧЕРИЗАЦИЯ В РЕЖИМЕ ПОЛЬЗОВАТЕЛЯ





HSA

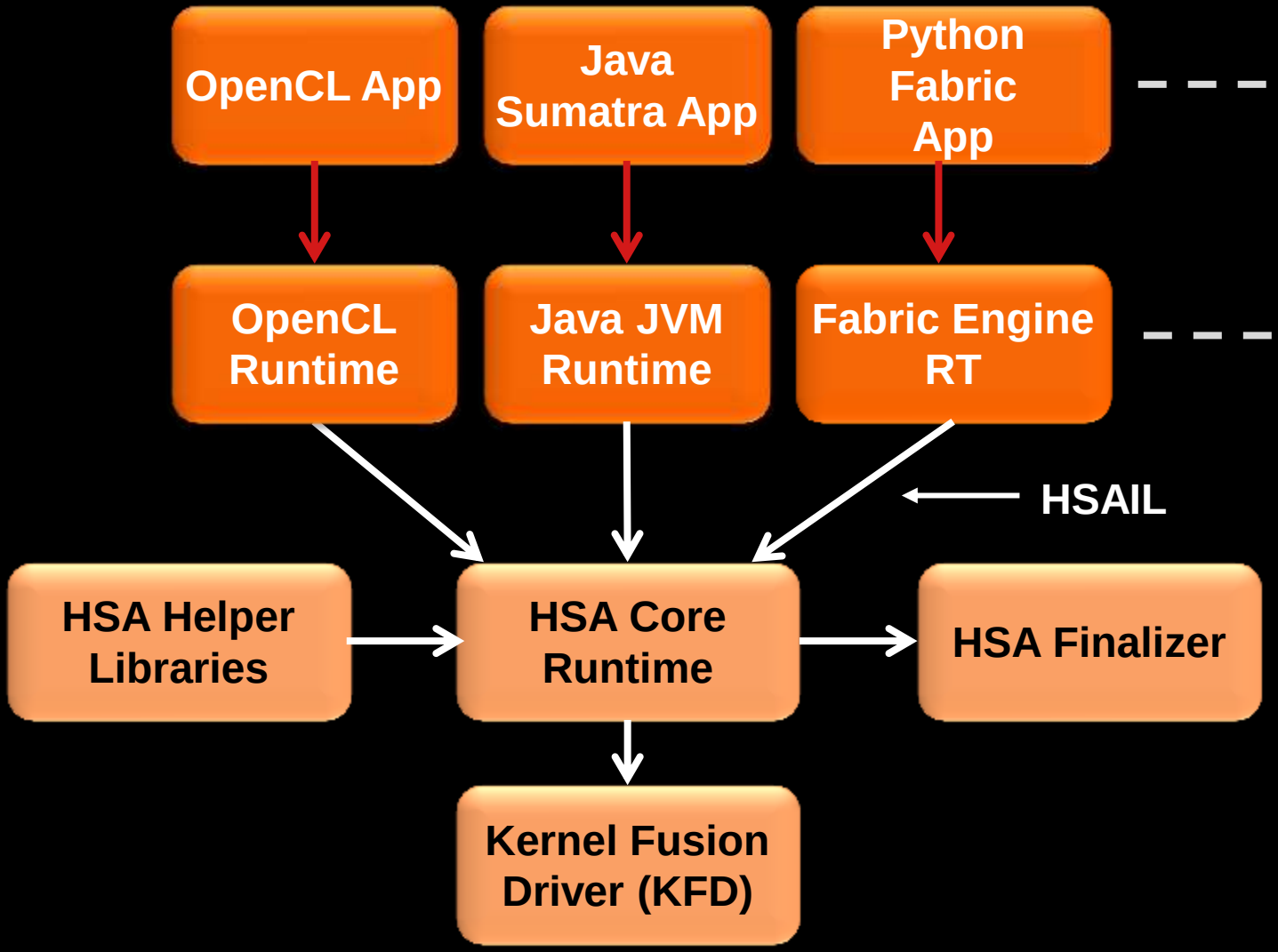

Application
OS
GPU
Queue Job
Start Job
Finish Job


YES FINALLY

ПОДДЕРЖКА ТЕХНОЛОГИЙ ПРОГРАММИРОВАНИЯ

C++ AMP
C++
C#
OpenCL
OpenMP
Java
Python
…

LANGUAGE SUPPORT ALLOWS HSA TO HAVE MULTIPLE SOFTWARE EXECUTION MODELS



JAVA ENABLEMENT BY APARAPI

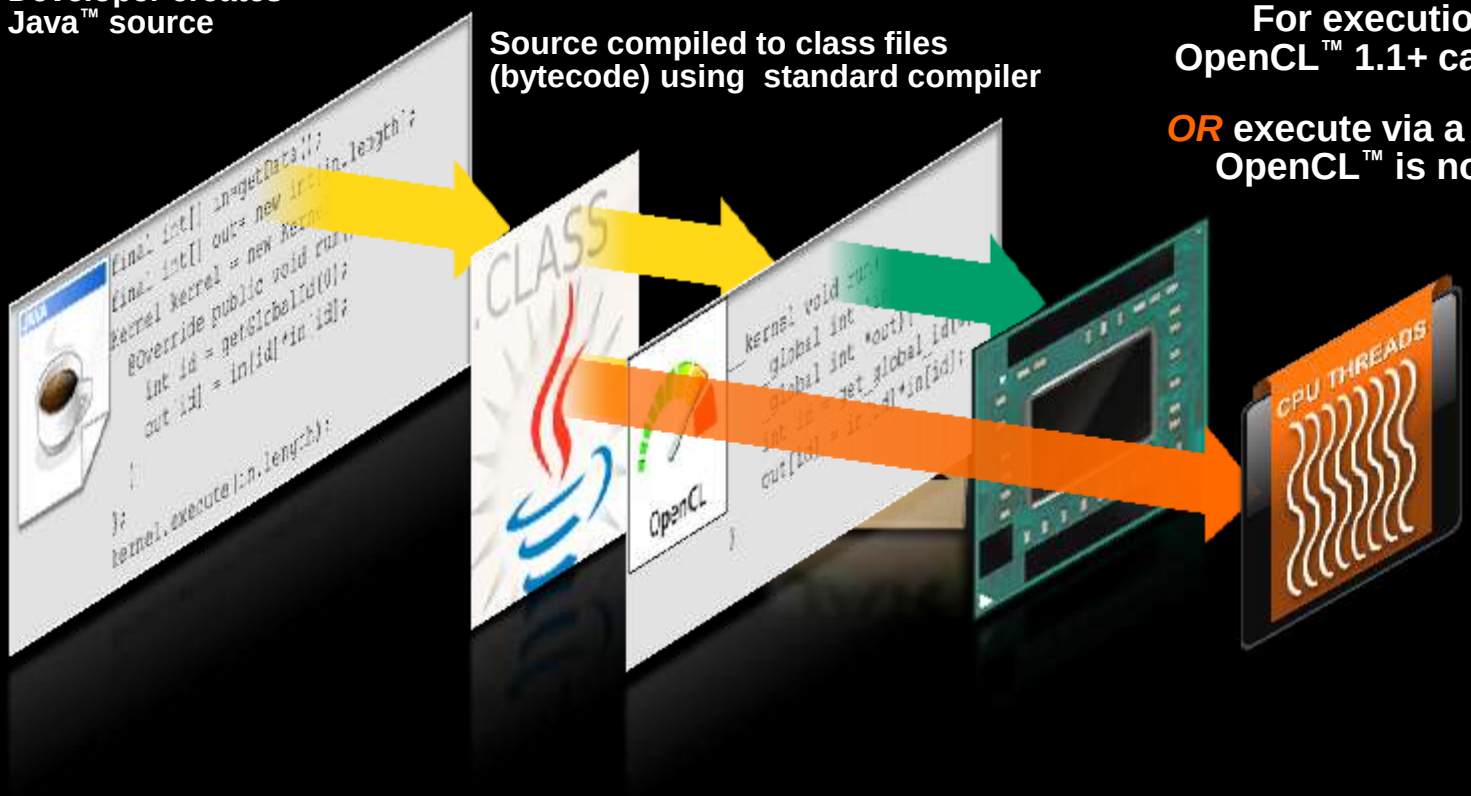
Aparapi = Runtime capable of converting Java™ bytecode to OpenCL™

Developer creates Java™ source

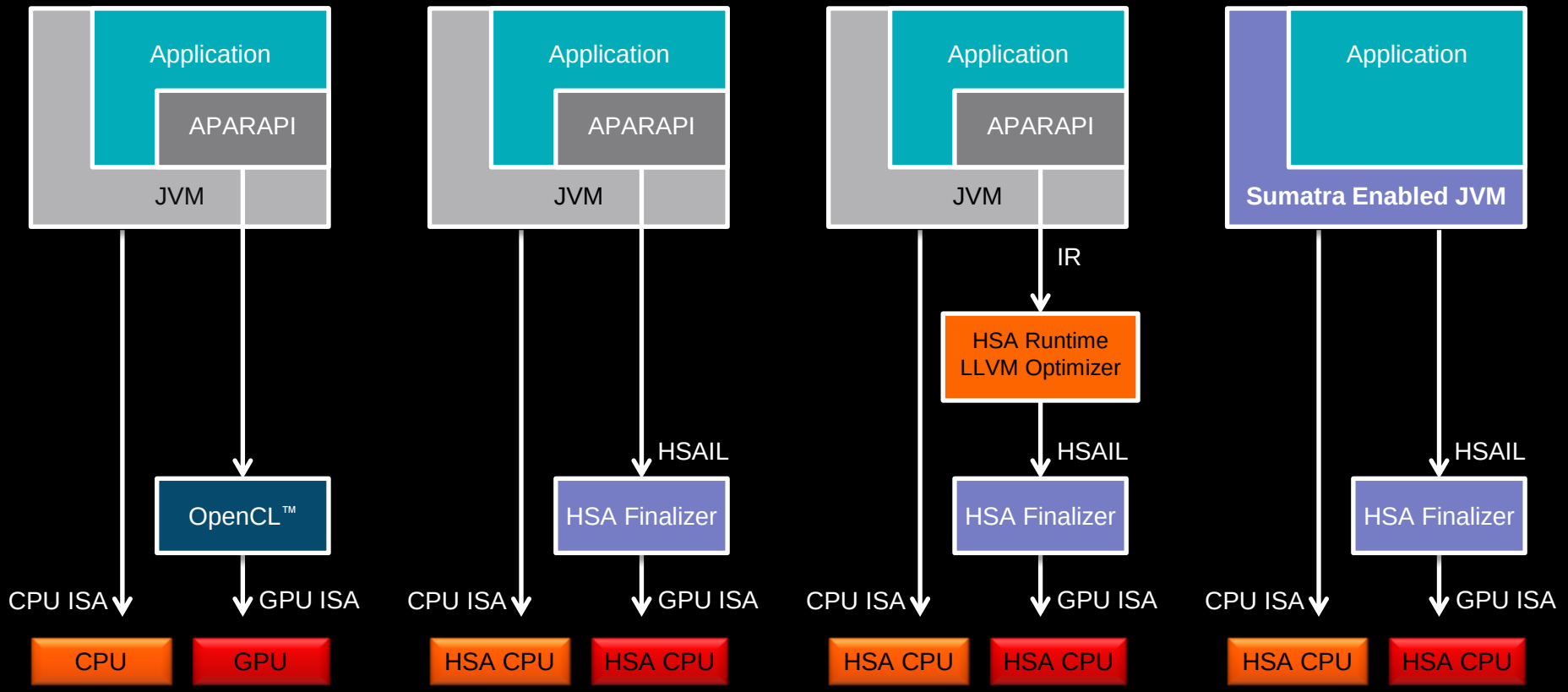
Source compiled to class files (bytecode) using standard compiler

For execution on any OpenCL™ 1.1+ capable device

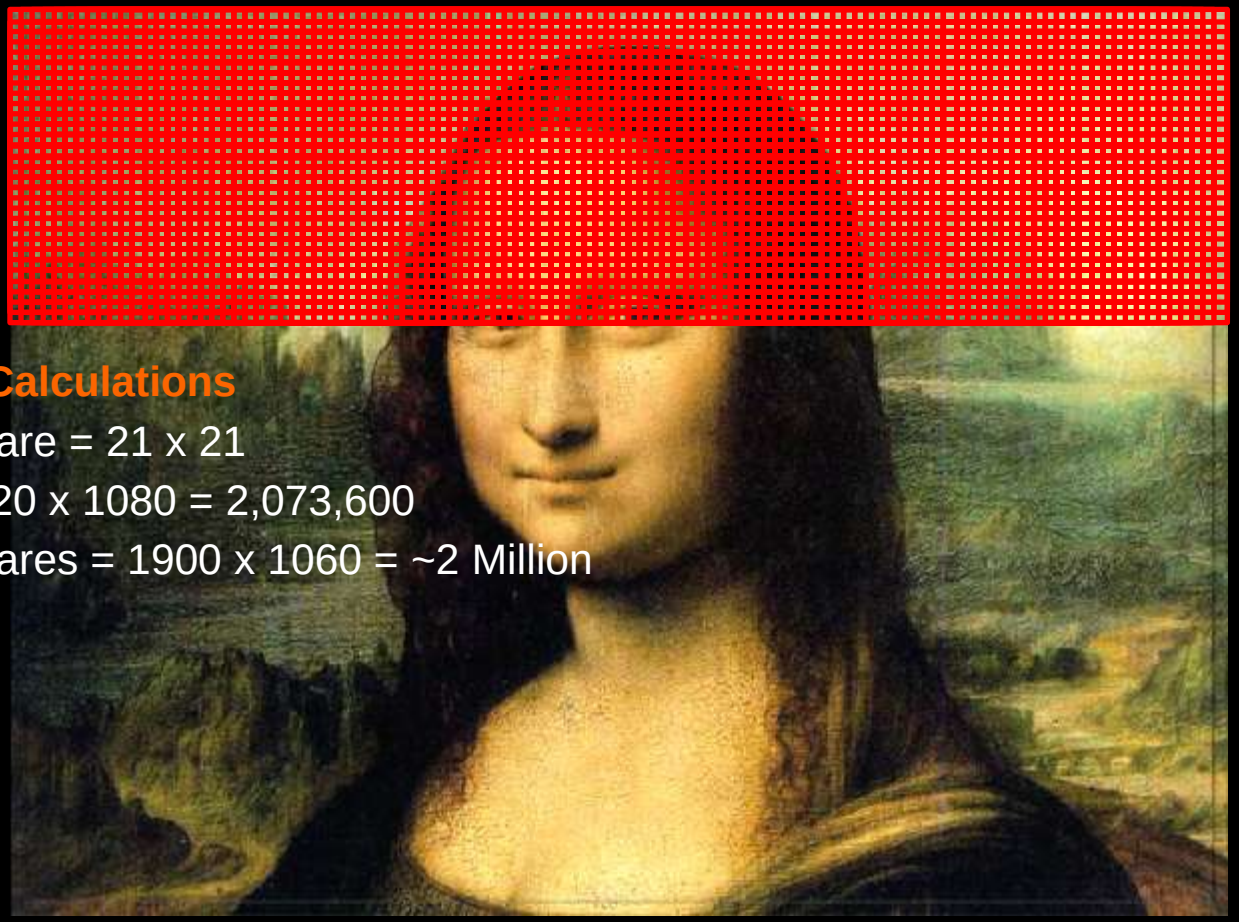
OR **execute via a thread pool if OpenCL™ is not available**



ПЛАНЫ ПО ПОДДЕРЖКЕ JAVA



LOOKING FOR FACES IN ALL THE RIGHT PLACES



Quick HD Calculations

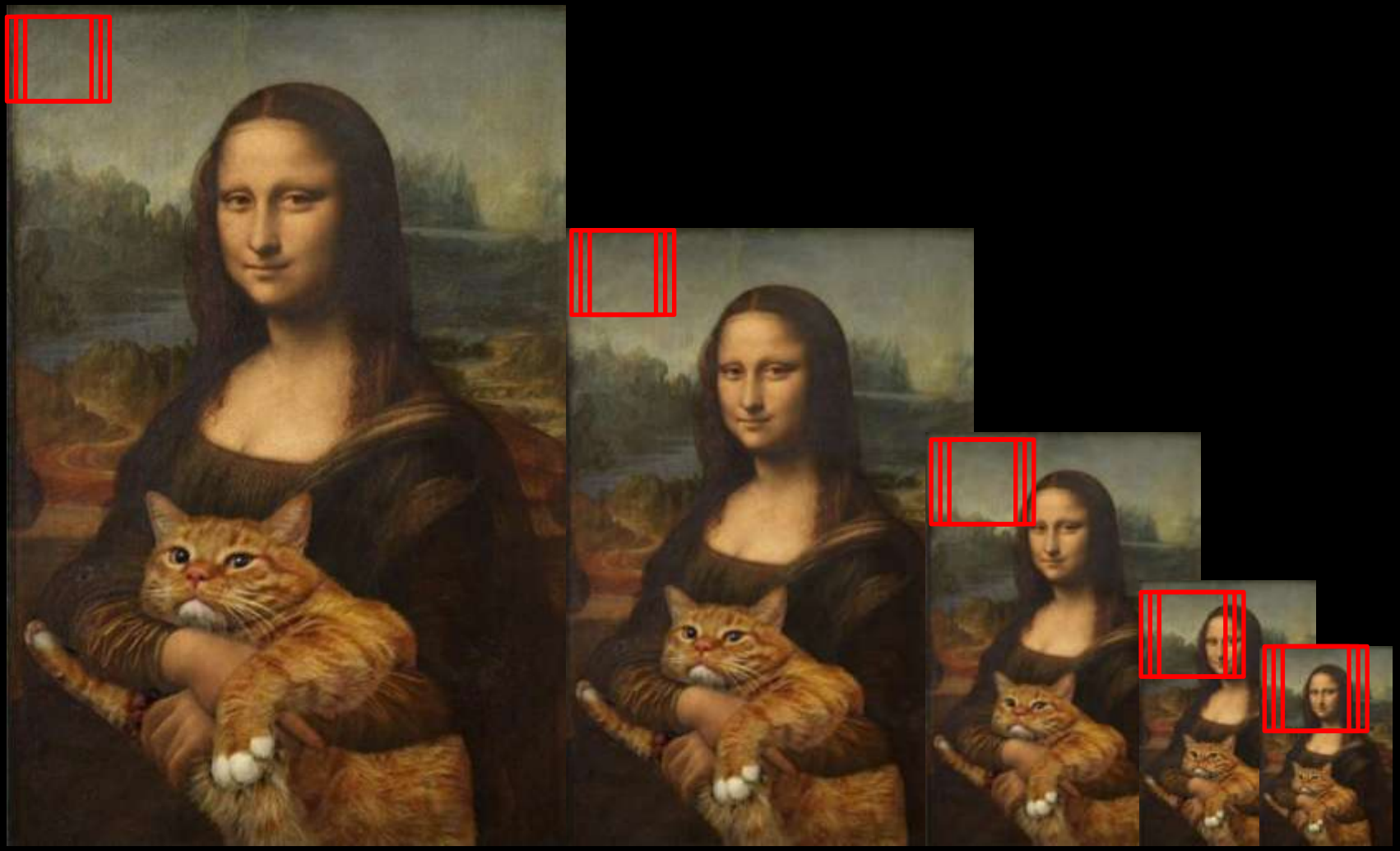
Search square = 21 x 21

Pixels = 1920 x 1080 = 2,073,600

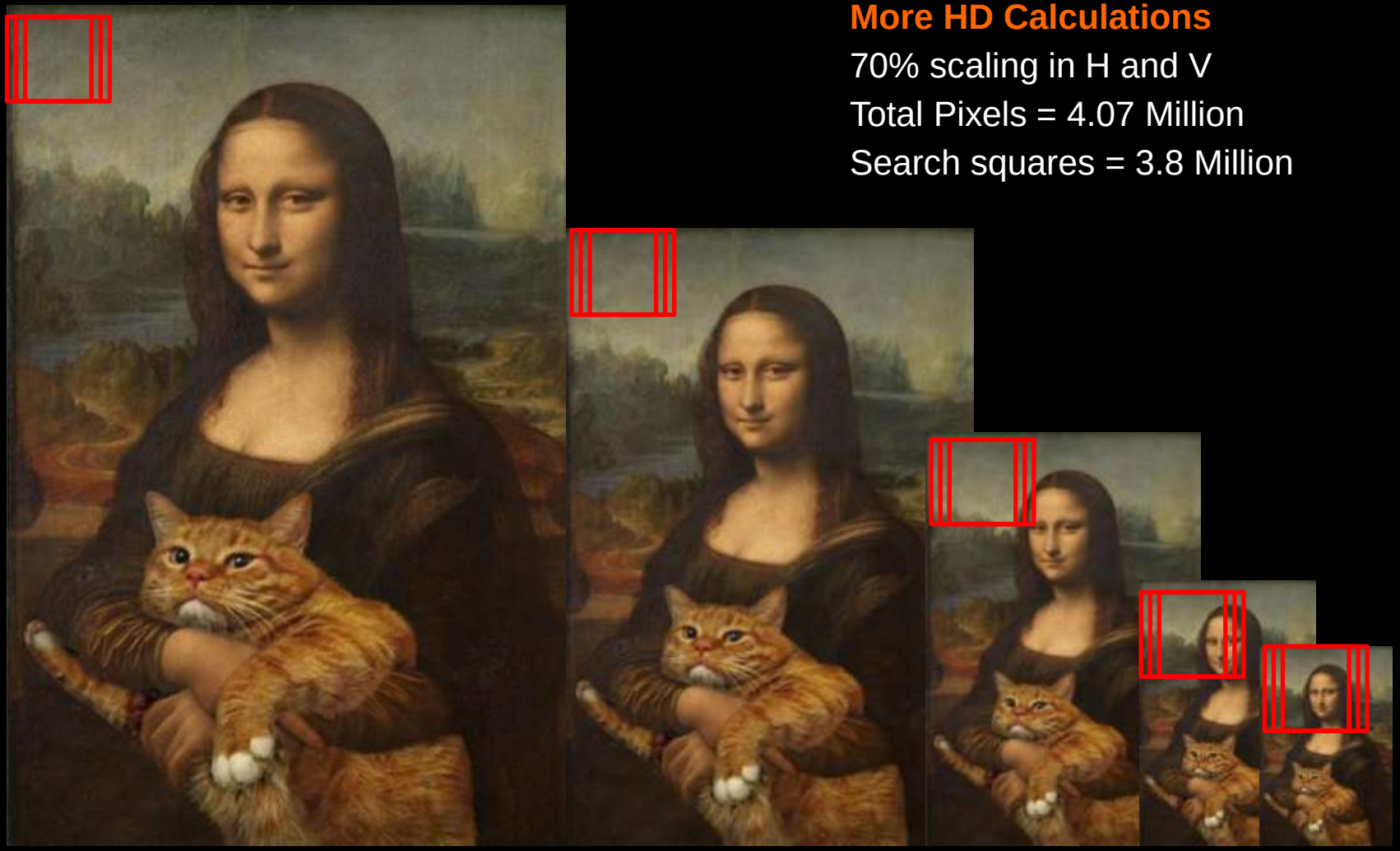
Search squares = 1900 x 1060 = ~2 Million

LOOKING FOR DIFFERENT SIZE FACES

BY SCALING THE VIDEO FRAME



LOOKING FOR DIFFERENT SIZE FACES
BY SCALING THE VIDEO FRAME



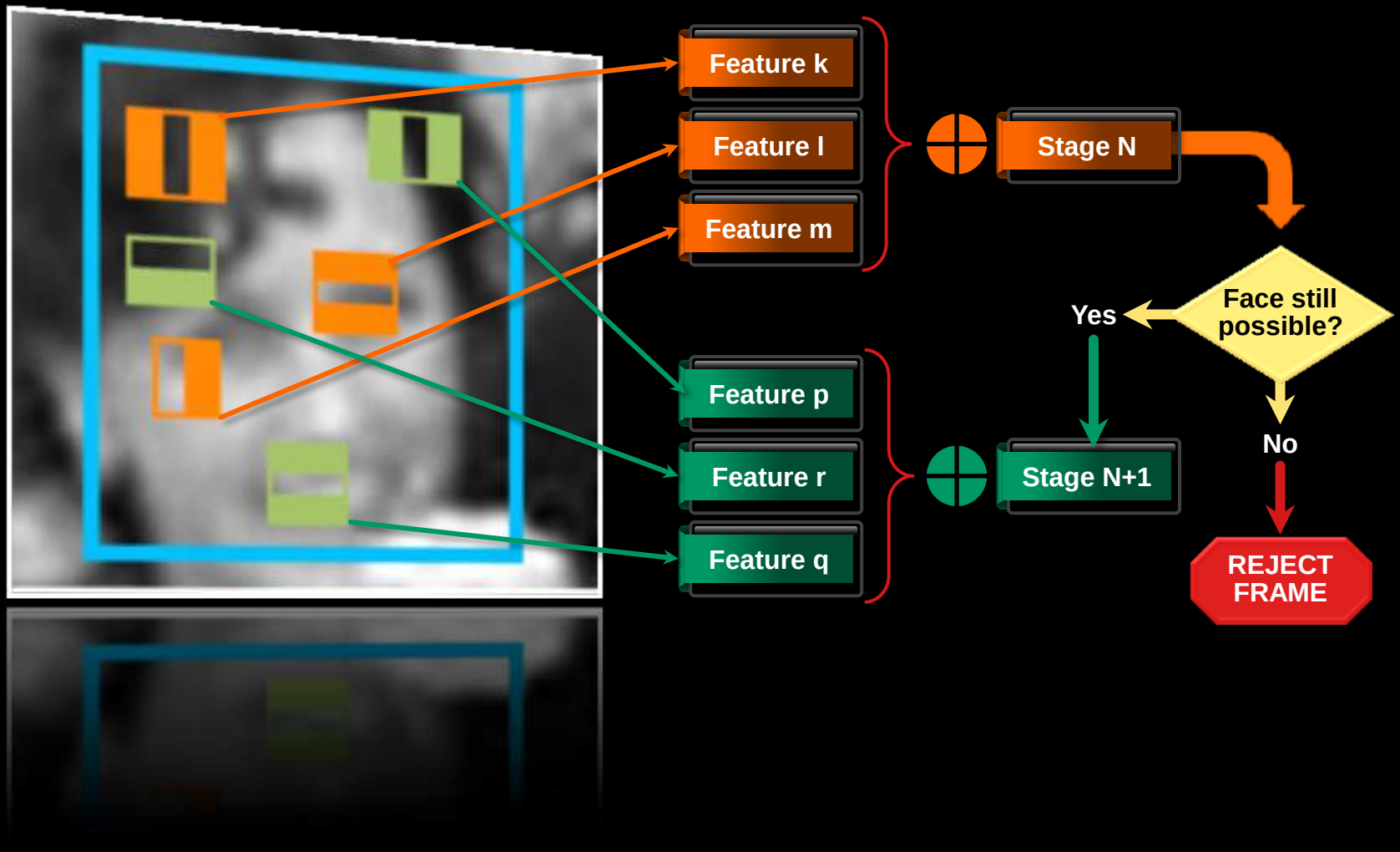
More HD Calculations

70% scaling in H and V

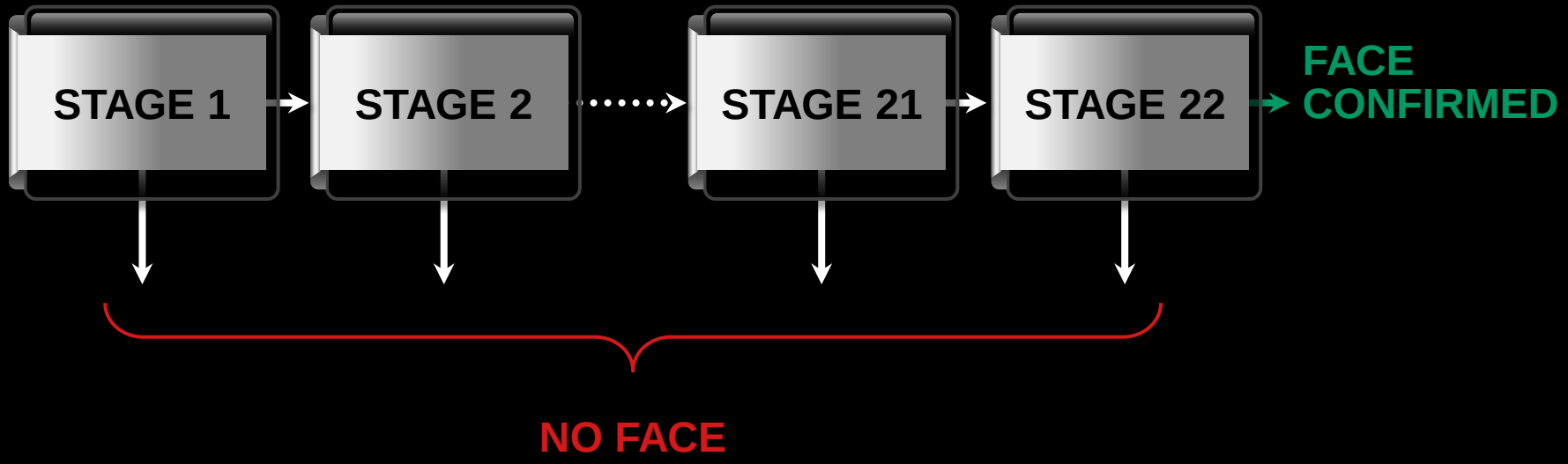
Total Pixels = 4.07 Million

Search squares = 3.8 Million

HAAR CASCADE STAGES



22 CASCADE STAGES, EARLY OUT BETWEEN EACH



Final HD Calculations

Search squares = 3.8 million

Average features per square = 124

Calculations per feature = 100

Calculations per frame = 47 GCalcs

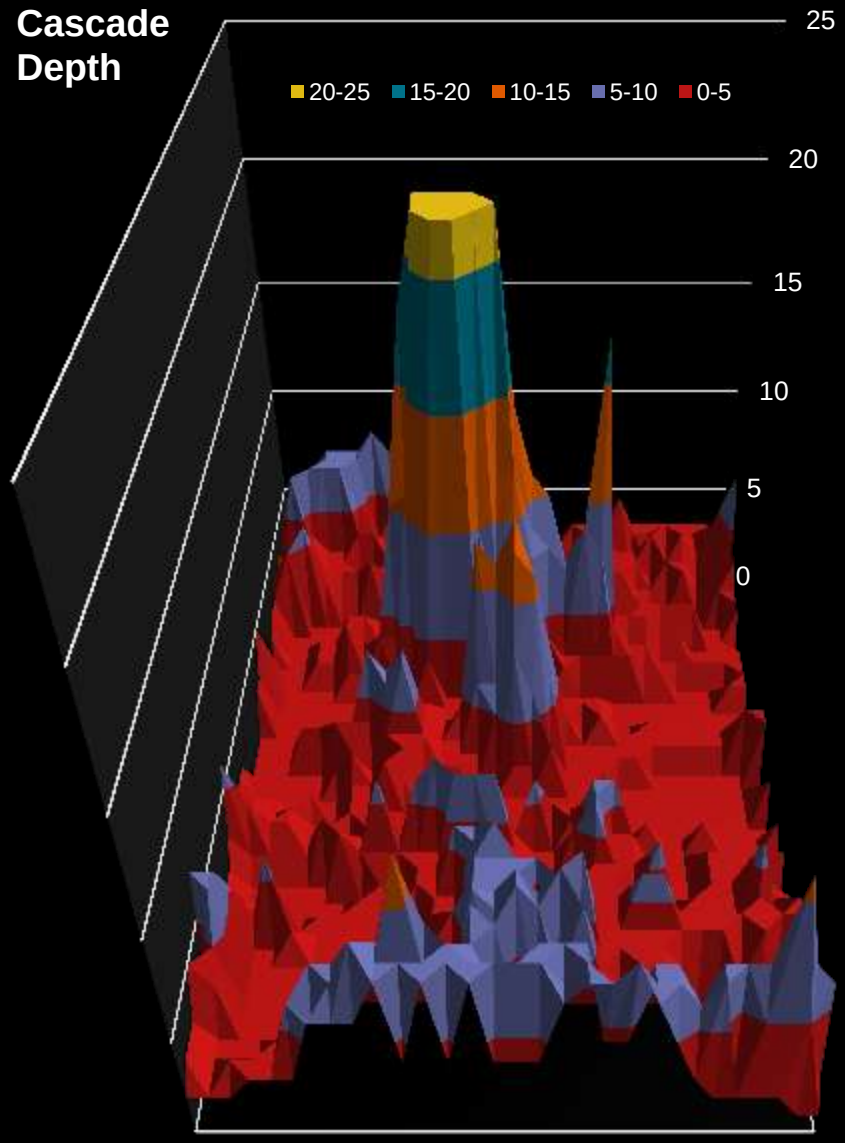
Calculation Rate

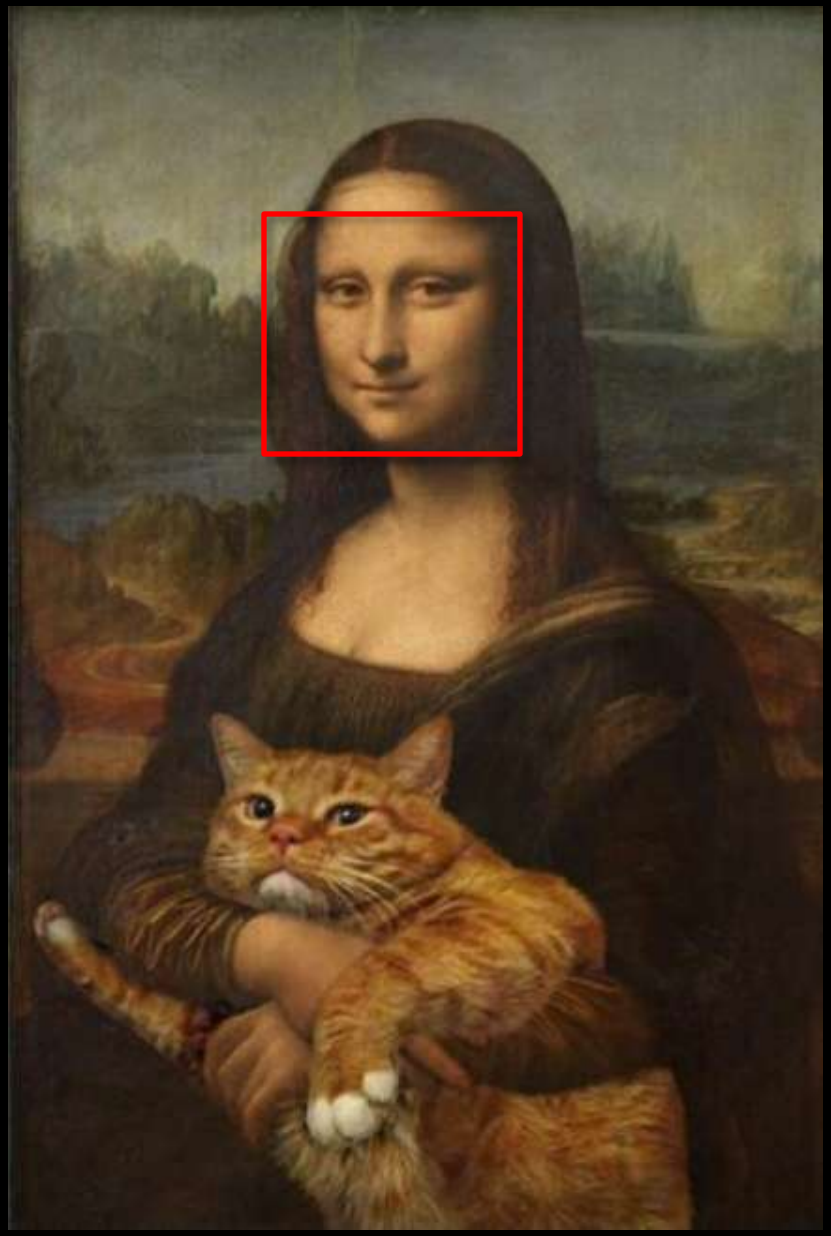
30 frames/sec = 1.4TCalcs/second

60 frames/sec = 2.8TCalcs/second

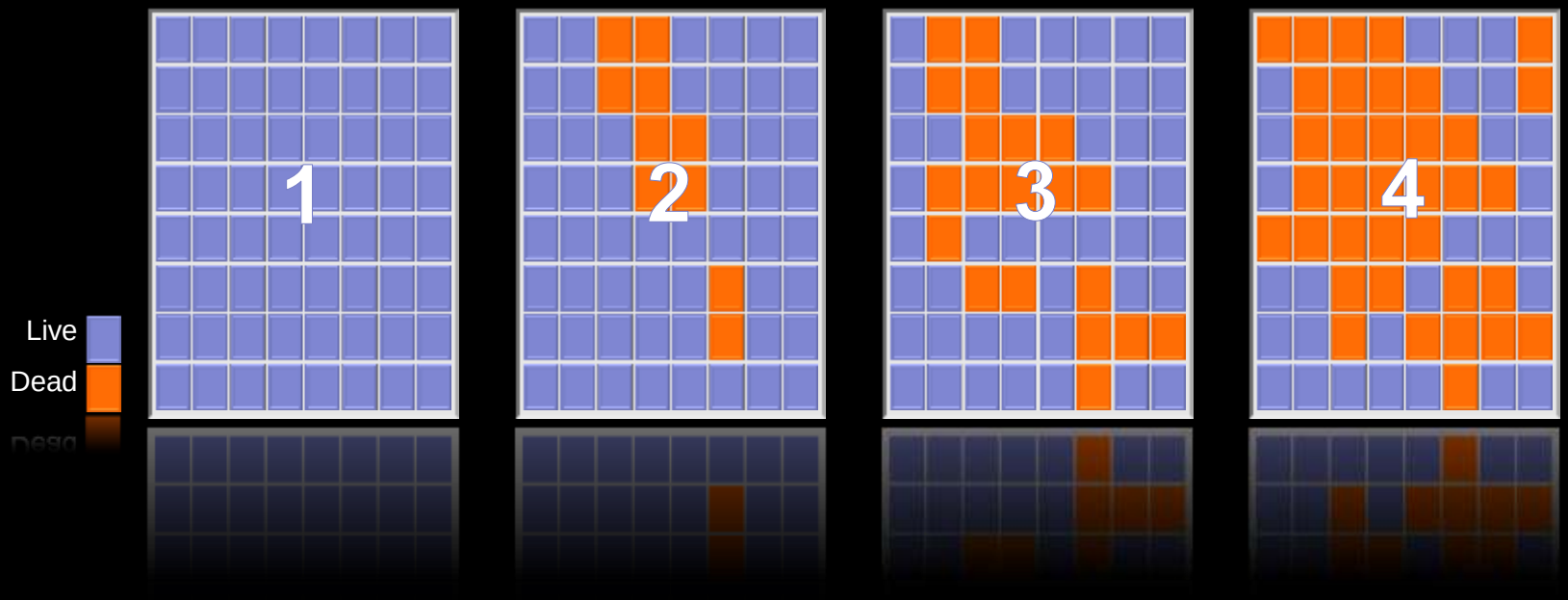
…and this only gets front-facing faces

CASCADE DEPTH ANALYSIS



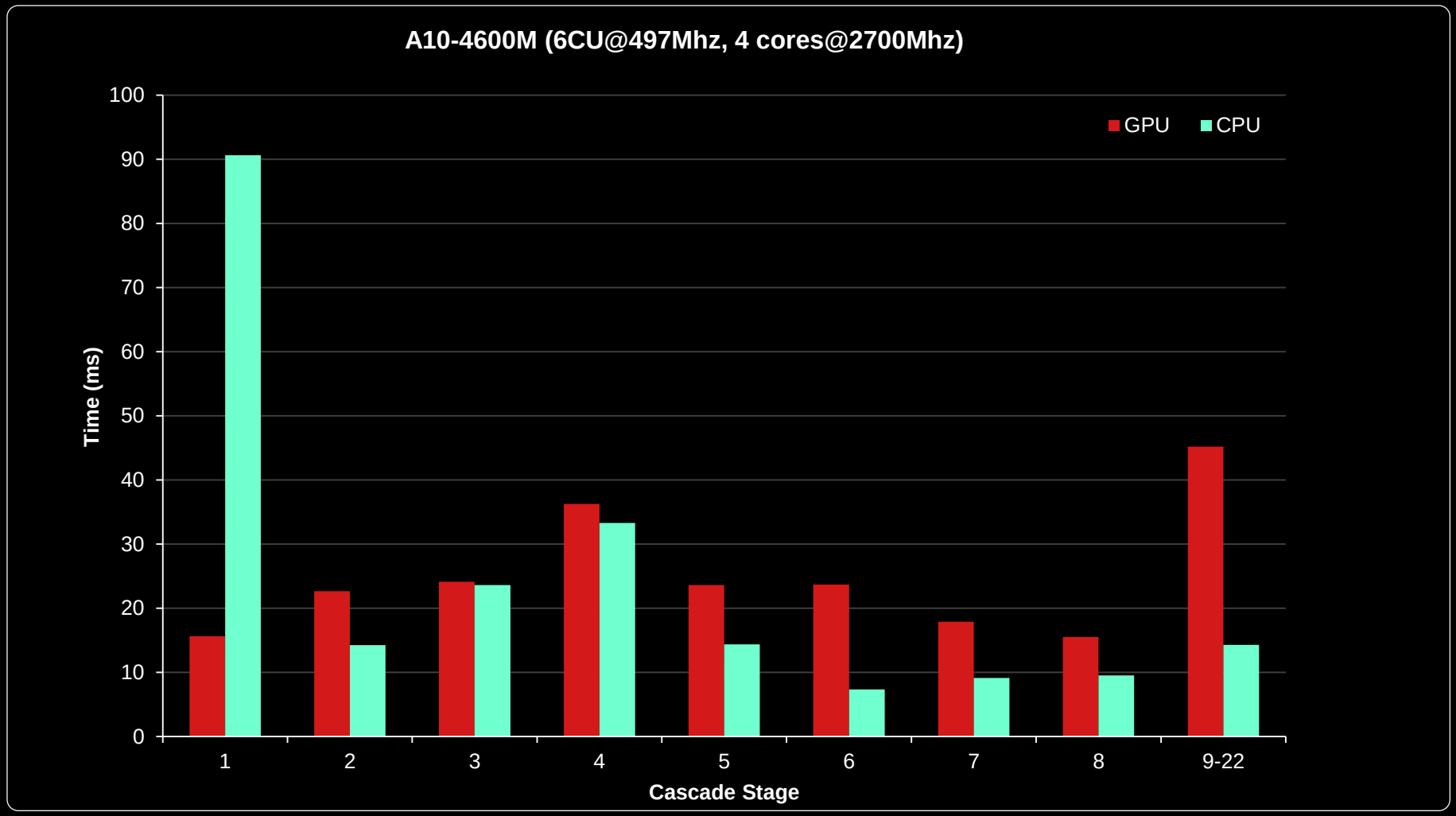


UNBALANCING DUE TO EXITS IN EARLIER CASCADE STAGES



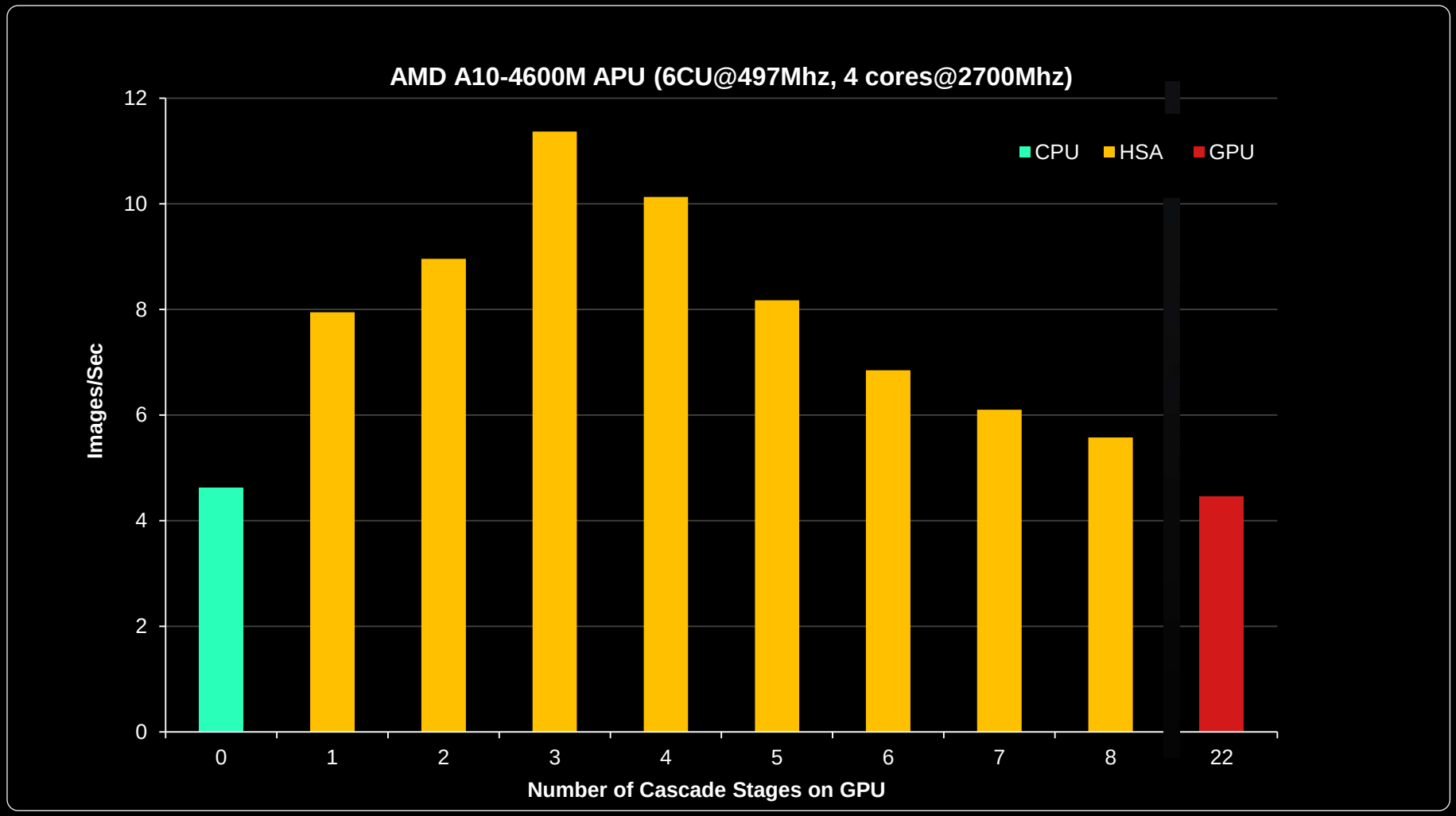
- When running on the GPU, we run each search rectangle on a separate work item
- Early out algorithms, like HAAR, exhibit divergence between work items
  - Some work items exit early
  - Their neighbors continue
  - SIMD packing suffers as a result

PROCESSING TIME/STAGE



AMD A10 4600M APU with Radeon™ HD Graphics; CPU: 4 cores @ 2.3 GHz (turbo 3.2 GHz); GPU: AMD Radeon HD 7660G, 6 compute units, 685MHz; 4GB RAM; Windows 7 (64-bit); OpenCL™ 1.1 (873.1)

ПРОИЗВОДИТЕЛЬНОСТЬ CPU-VS-GPU

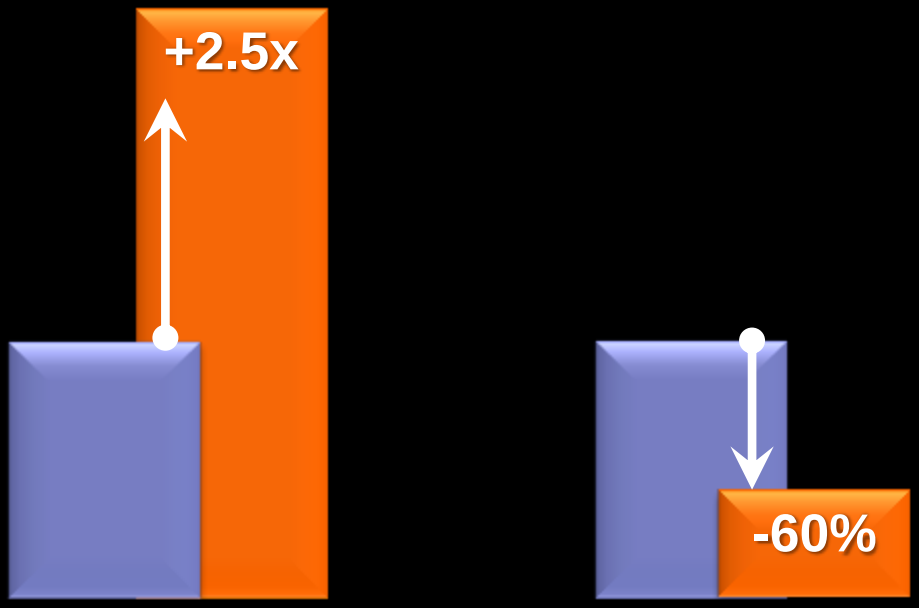


AMD A10 4600M APU with Radeon™ HD Graphics; CPU: 4 cores @ 2.3 MHz (turbo 3.2 GHz); GPU: AMD Radeon HD 7660G, 6 compute units, 685MHz; 4GB RAM; Windows 7 (64-bit); OpenCL™ 1.1 (873.1)

HAAR SOLUTION

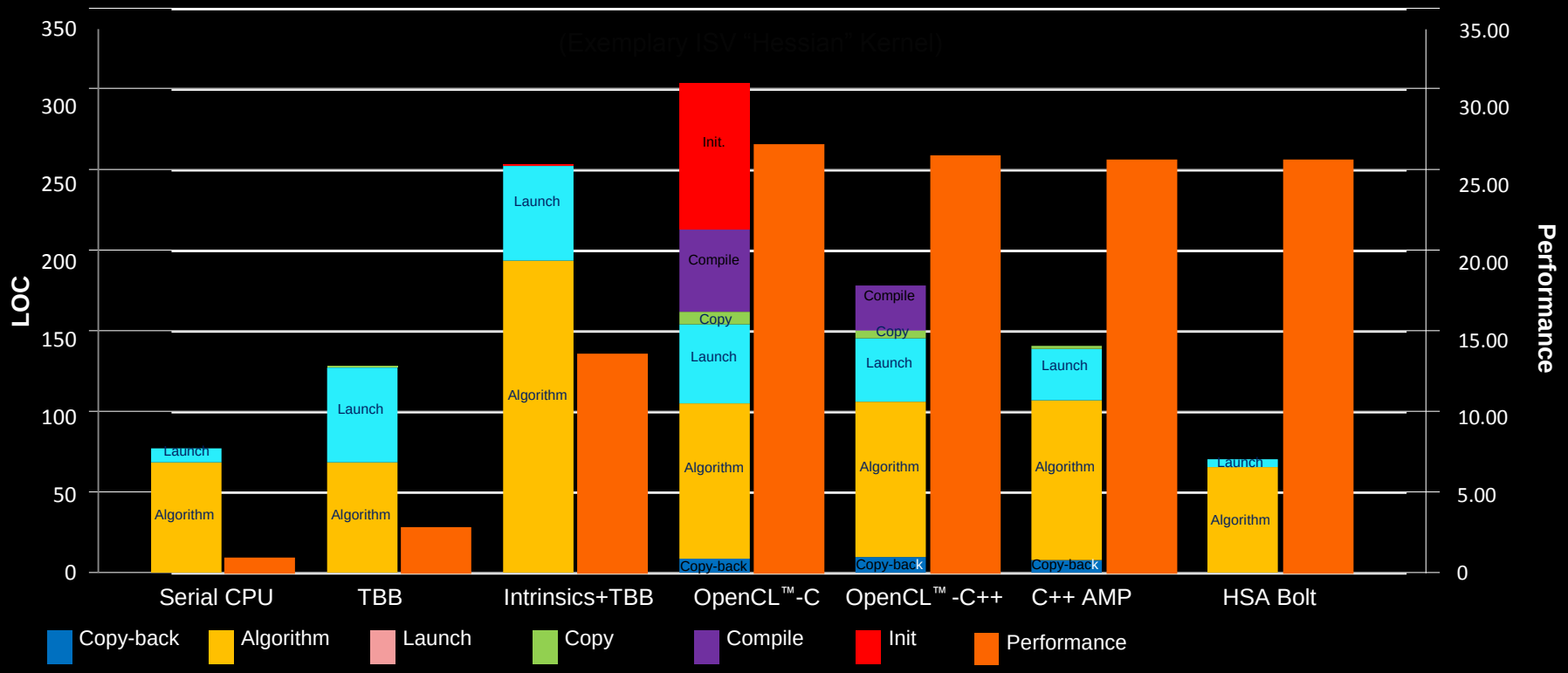
RUN DIFFERENT CASCADES ON GPU AND CPU

By seamlessly sharing data between CPU and GPU,
HSA allows the right processor to handle its appropriate workload



LINES-OF-CODE AND PERFORMANCE FOR DIFFERENT PROGRAMMING MODELS

("Hessian" kernel)



AMD A10-5800K APU with Radeon™ HD Graphics – CPU: 4 cores, 3800MHz (4200MHz Turbo); GPU: AMD Radeon HD 7660D, 6 compute units, 800MHz; 4GB RAM.
Software – Windows 7 Professional SP1 (64-bit OS); AMD OpenCL™ 1.2 AMD-APP (937.2); Microsoft Visual Studio 11 Beta

MORE INFO AT

http://hsafoundation.com

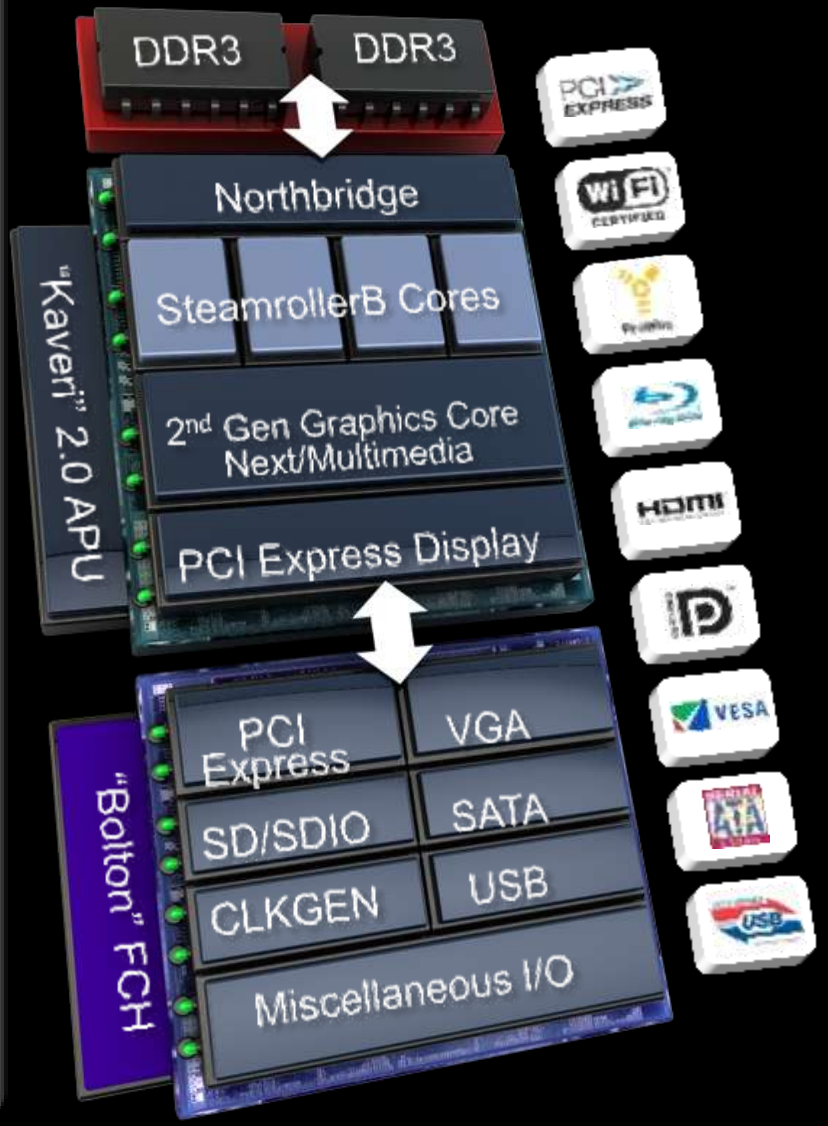
HSA РЕШЕНИЯ ОТ ARM (CORTEX-A15 MALI-T600)

- Google Chromebook
- Google Nexus 10
- InSignal Arndale Community Board

AMD “KAVERI” 2.0 PLATFORM DETAILS

APU Features

- ***New*** “SteamrollerB” CPU Core with up to 20% performance* increase over “Richland”
  - Up to 4 Steamroller cores and 4 MB total L2 cache
  - Temperature Smart Turbo Core
- ***New*** Power Optimized Graphics Core Next with up to 30% performance* increase over “Richland”
  - Multiple DirectX® 11.1 GPU configurations
  - Dual Graphics support with “Crystal” Series
- ***New*** AMD fixed function acceleration
  - UVD 4.2 Universal Video Decode Engine
  - VCE 2.0 Video Compression Engine
  - ***New*** ACP (Audio Co-Processor)
  - SAMU 2.1 Secure Asset Management Unit
- ***New*** Display and I/O Features
  - ***New*** PCIe Gen3 x16 for discrete GPU expansion
  - ***New*** Dedicated PCIe SSD interface
  - PCIe Gen2 1 x4, 4 x1, 1x4 UMI
  - 4096 x 2160 resolution per display output
  - 16K x 16K Max Eyefinity SLS resolution
  - “Lightning Bolt” Docking Solution
- Power Management and Battery Life
  - 35W to 15W TDPs
  - Targeting ~11 Hours Battery Life MM07* 62WHr (~8.5Hr 45WHr) 14” 1366x768 eDP panel
  - ***New*** AMD Start Now 3.0 with smart sleep



*Pre-silicon projection and not based on actual measured data

QUESTIONS AND ANSWERS