Heterogeneous Computing in ARM Architecture

Media Processing Division

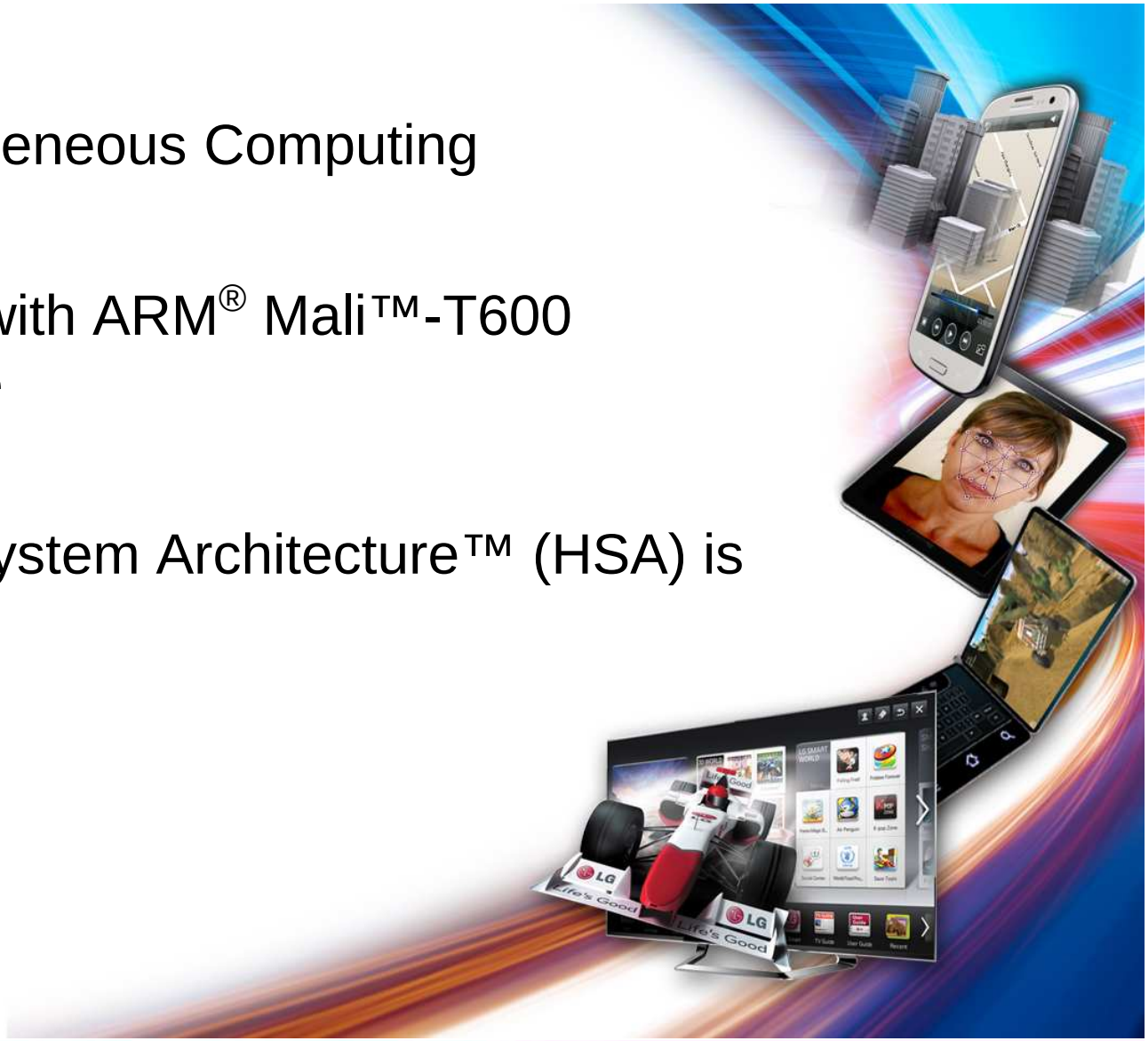
ARM

June 25th 2013

ARM

Agenda

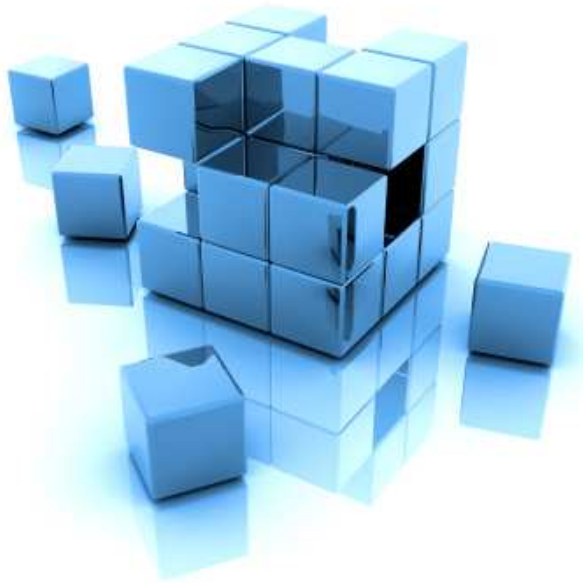
- Trends in Heterogeneous Computing
- GPU Computing with ARM® Mali™-T600 series as example
- Heterogeneous System Architecture™ (HSA) is the future



ARM and HSA

Trends in the Industry

- Heterogeneous multiprocessing
  - Established approach for SoC design
  - Mix of many specialized accelerators, implementing different ISAs
  - Diverse programming approaches lead to lack of portability
- Parallel computation for performance and efficiency
  - Endorsed at all levels of computer architecture
  - Parallel programming traditionally difficult
- General purpose programmability of GPUs
  - Massive parallel computation potential
  - Increasing programmability

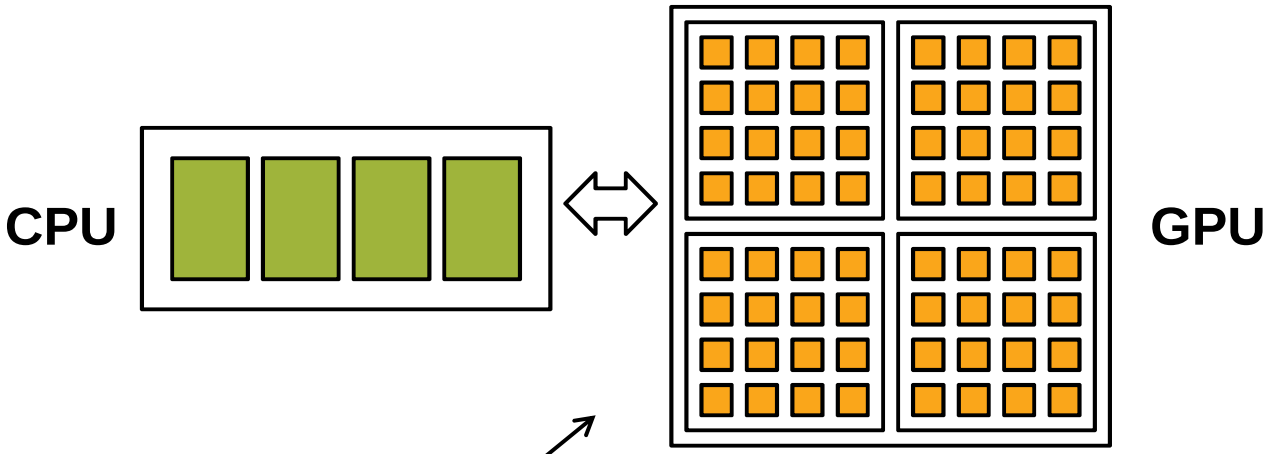


What is Parallel Computing?

- Simply, doing multiple tasks simultaneously
- **Task-Parallel computing** does different tasks concurrently
  - Reading email, playing music, and surfing the web are all separate tasks
  - In a multicore system, these can execute simultaneously
- **Data-Parallel computing** does the same operation on a collection of data concurrently
  - Adjusting the contrast of the pixels of an image
  - Each thread executes the same code but with different data
    - Classic SIMD (single-instruction, multiple-data)
- GPU computing is perfect for data-parallel applications

What is Heterogeneous Computing?

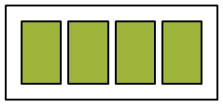
GPU used as computational accelerators or companion processors



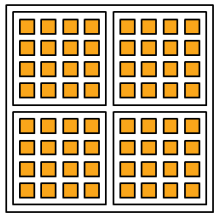
Massively parallel architecture gives great computational capabilities

Cost effective, efficient, great floating point performance

Complementary Processor Architectures

The CPU 

- Serial workloads and task parallel workloads
- < 10 threads
- 1-4 cores
- Short pipeline, <20 stages
- Low latency
- General purpose
- SIMD engine

The GPU 

- Data parallel workloads
- 100s-1000s threads
- 1-100s cores
- Long pipeline, >50 stages
- Very high latency
- High throughput
- 2D/3D Graphics
- Stream processing

GPU Compute Making the Difference

Trends

- Heterogeneous computing
- Portability
- Parallel computation
- Hardware acceleration
- GPU computing

Benefits

- More efficient processing
- Improved accuracy/quality
- BOM reduction
- Unlock new use cases
- Improved existing use cases

2D to 3D

Multi-Perspective Vision

Computer Vision

Computational Photography

Multi-User Interaction

Information Extraction

Real Time Still and Moving Image Perfection

Light-Field Photography

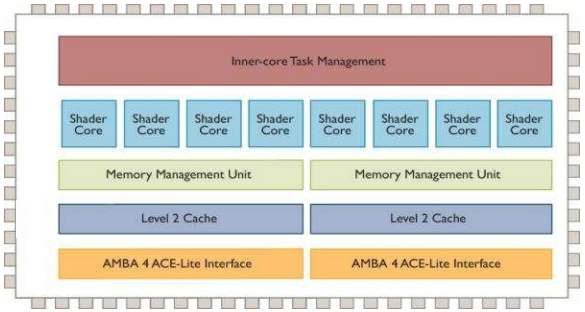
Up scaling

GPU COMPUTING
Mali-T600 as Example

Mali-T600 GPU Series Overview

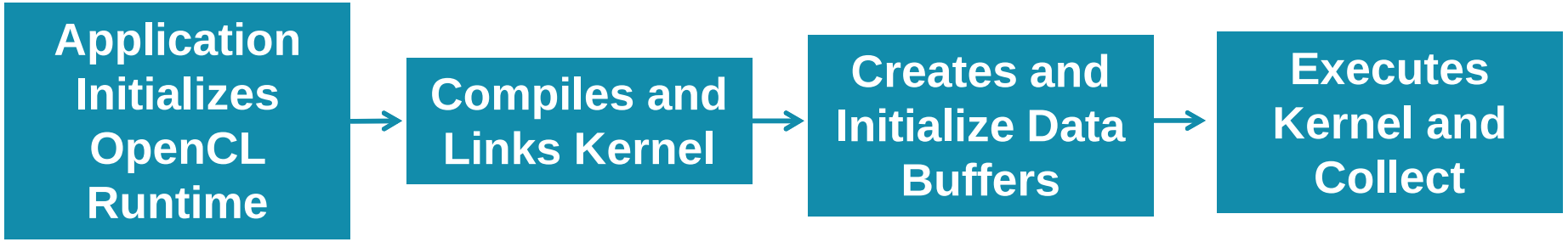
- Innovation and market leadership
  - Tri-pipe ALU design - optimal graphics and GPU compute
  - Native 64-bit integer and floating point (IEEE 754-2008), scalar and SIMD
- Flexibility and scalability
  - Mali-T624 and Mali-T628 for smartphones and SmartTVs
  - Mali-T678 for the best in compute and graphics for tablets
- Software compatibility and comprehensive API support
  - DirectX® 11, OpenGL® ES 3.0
  - OpenCL™ Full Profile and Renderscript™ compute
- Performance
  - 100s of GFLOPs of arithmetic performance

Mali-T628



What about OpenCL?

- OpenCL is an API for heterogeneous computing
  - Write one source, deploy on many type of processors
- Currently, it's targeted for data-parallel applications
- Applications use kernels to process data provided to the OpenCL runtime
  - Kernels are written in OpenCL C
    - Subset of C99 with the addition of vector data types (e.g. float4)



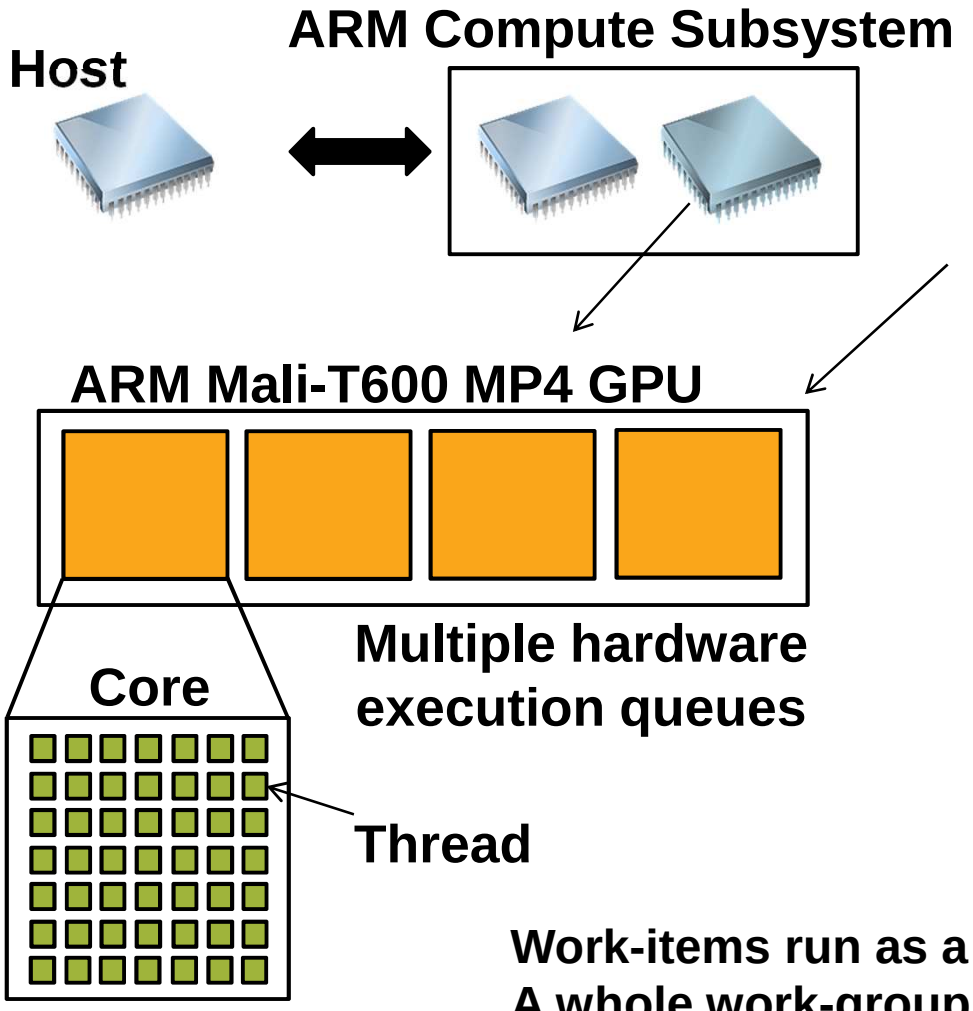
GPU Computing with no compromises

- Embedded Profile is a subset of Full Profile, reducing features and precision
- All shipping processors openly programmable with OpenCL 1.1 are Full Profile
- All mainstream developers are producing for Full Profile
- All existing software in the industry has been developed for Full Profile

Feature	Benefit
Native support for 64-bit integer maths (scalar and SIMD)	Radically faster and more efficient than software emulation Beneficial for multimedia encoders/decoders and encryption software, pointer arithmetic for the post 4Gb world, large counters
IEEE 754-2008 compliance	Same floating point accuracy on a Mali-T600 Series GPU as any other Full Profile conformant platform
Hardware accelerated support for 3D images	Great for volumetric modelling Useful in physics, games
Built-in atomic operations	Accelerated in hardware on Mali-T600 No need for expensive external memory synchronization or emulation Cornerstone of parallel computation

With Mali-T600, ARM is the first IP vendor to pass conformance for OpenCL 1.1 Full Profile

OpenCL Platform Model on Mali-T600

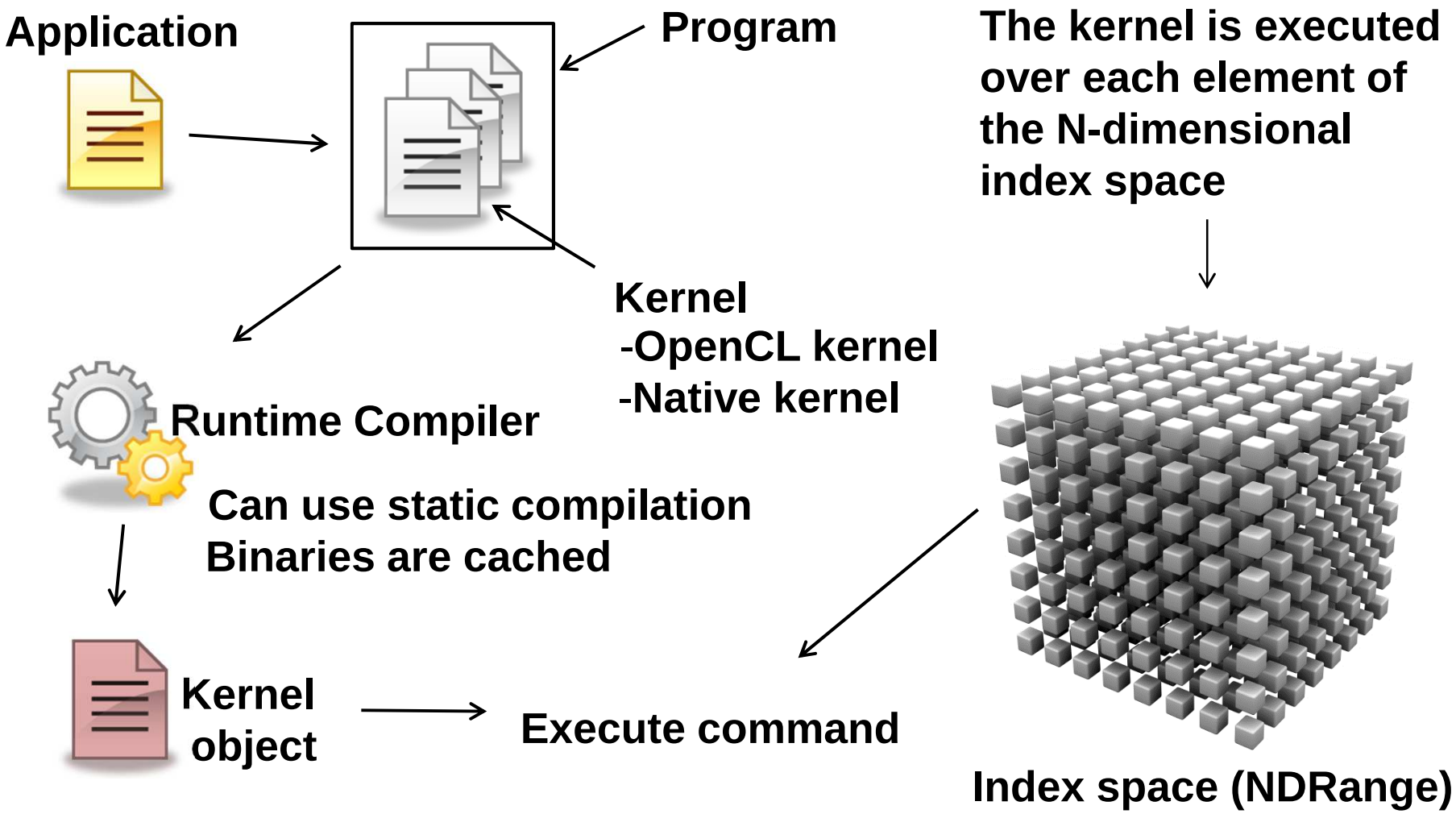


Job manager handles everything in hardware:

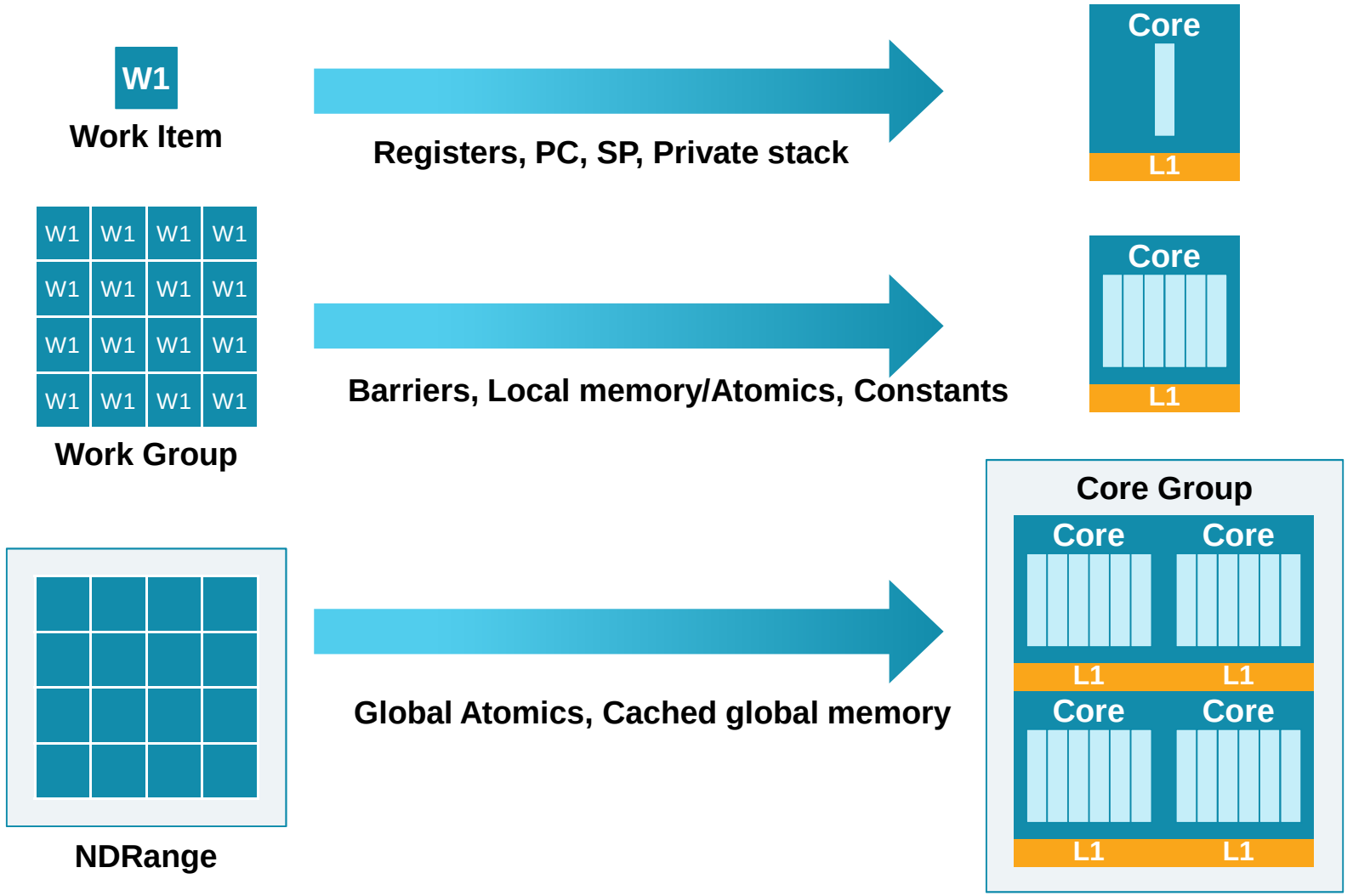
- Issuing all tasks to available cores
- Handling out-of-order execution queues
- Continually spawning work items (threads) to keep cores busy
- Providing work item IDs
- Per-job completion interrupts can be requested

Work-items run as a single thread on a core
A whole work-group executes on a single core
Each thread has its own registers, PS, SP, private stack

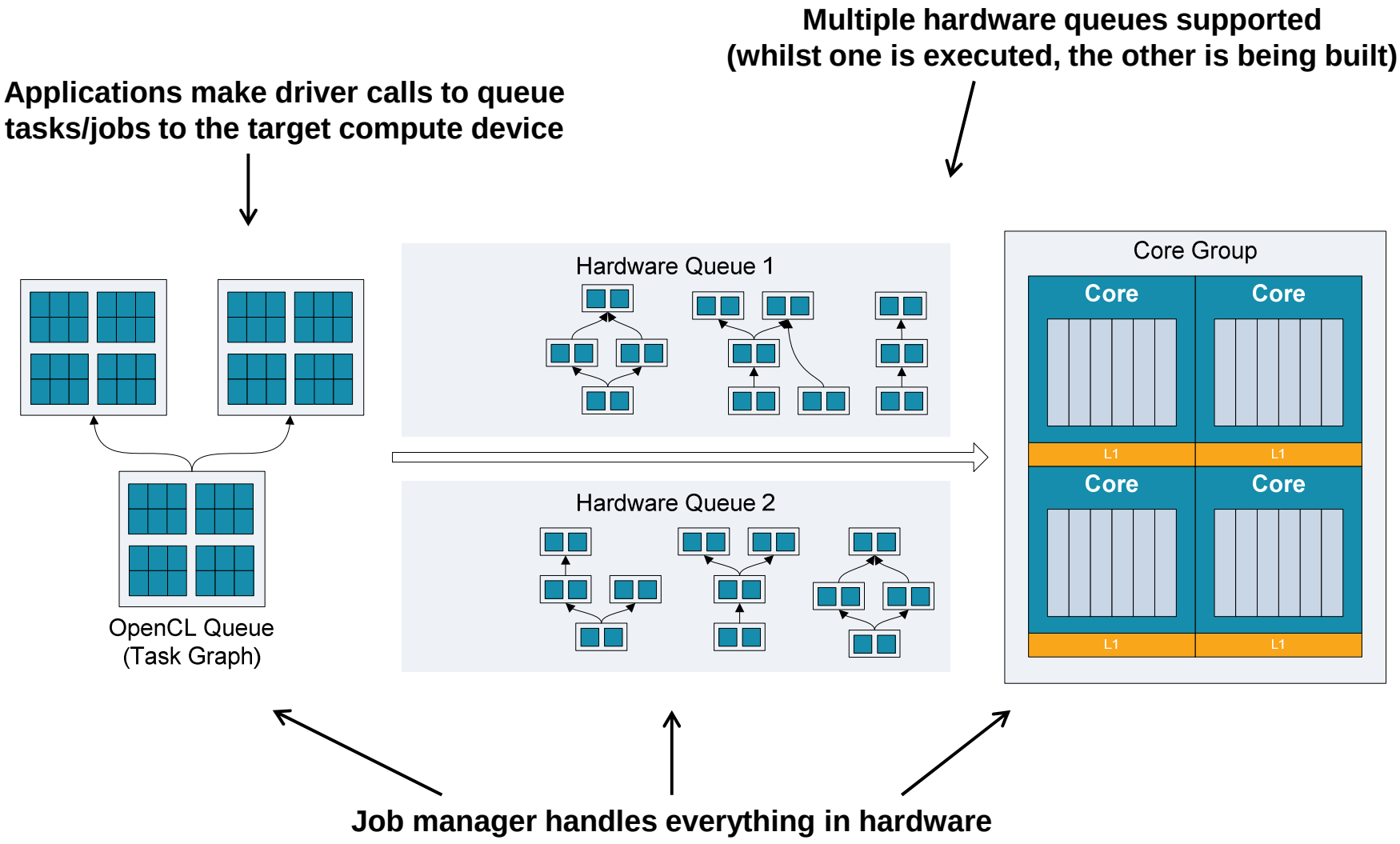
OpenCL Programming Model



OpenCL Execution Model on Mali-T600

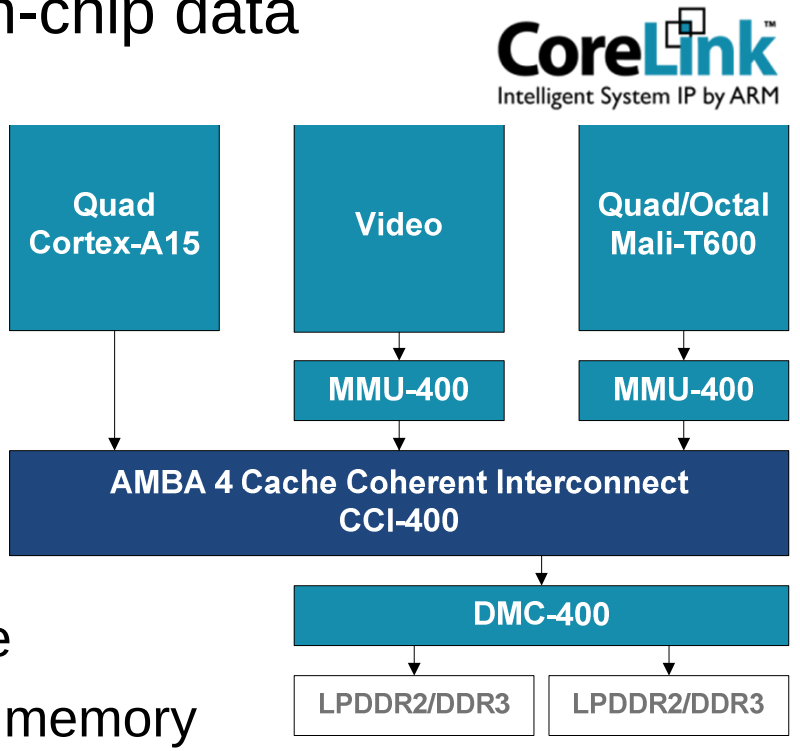


OpenCL Execution Model on Mali-T600



Coherency on Cortex™-A15 & Mali-T600

- Coherency allows the sharing of on-chip data
  - Reduces external memory access
  - Saves power
- Compute subsystems for SoC
  - Designed and optimized by ARM
- Cache Coherent Interconnect
  - Enables hardware cache coherency
  - Increases available CPU performance
  - Reduces the need to access external memory



- Improved OpenCL performance across CPU and GPU
  - GPU snoops into CPU caches
  - Enables simple sharing of data between processors

GPU COMPUTING ON MALI

Mali GPU Compute is here now!

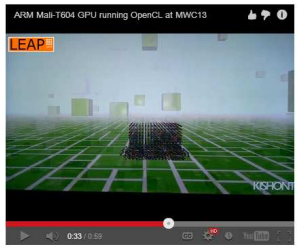
- **Certified Khronos Conformant**
  - OpenCL 1.1 Full Profile on Linux and Android
- **Mature, Proven in Silicon**
  - Samsung Exynos 5 Dual, implements Full Profile
  - OpenCL and Renderscript DDK available now
  - Proven performance benefits with Kishonti Benchmarks
- **Shipping in real products**
  - Google Chromebook
  - Google Nexus 10
  - InSignal Arndale Community Board
- **API exposed for developers**
  - OpenCL on Linux for Arndale platform
  - Renderscript computation on Android for Nexus 10

Nexus 10
Google Experience Device for Android 4.2
Pioneering GPU Computing on Android

Chromebook

Arndale
First OpenCL low-cost embedded dev platformd

Mali GPU Computing Demos



KISHONTI INFORMATICS

MULTICORE WARE

Physics Simulation (ARM)

Morpho

OpenCV Face Detection (ARM)

Aptina IMAGING

Synthesis Corporation

apical

Ittiam

http://goo.gl/rE61Q

Advanced Image Processing

MULTICORE WARE

- RenderScript is the official Heterogeneous Compute Android API
- Since Android ICS 4.2 it has been enabled to target the GPU
- Complex image filters can be greatly accelerated by GPU Compute

Filter	Speed-up [1]
MotionBlur	3.5x
Cloud	4.2x
Labyrinth	3.8x
TitleReflection	7.3x
WhirlPinch	3.6x
Wave	7.0x
Bicubic	15.4x

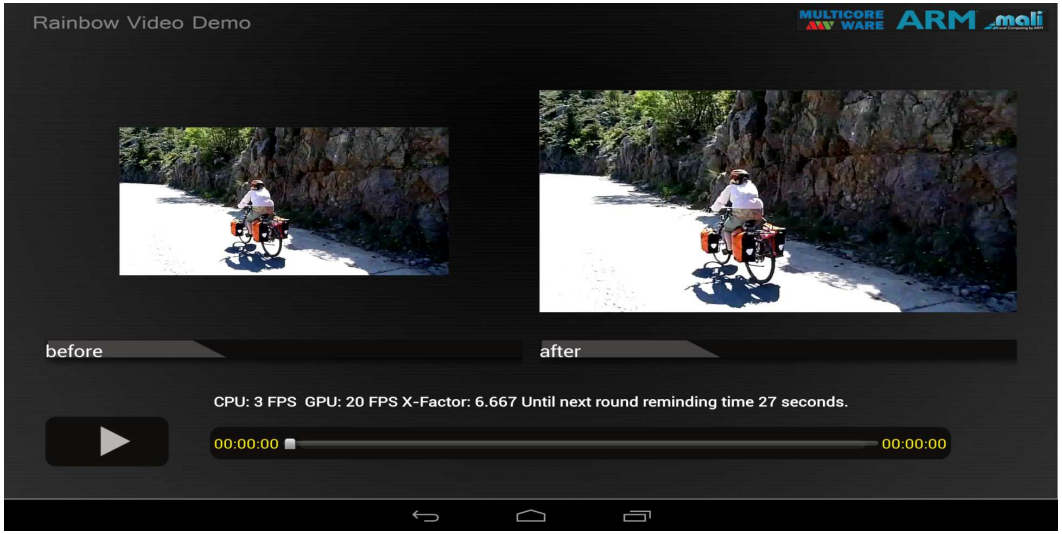
Image sise: 2560x1920

[1] Acceleration compares RenderScript compiled on device (LLVM) on dual-core Cortex™-A15 and Mali™-T604 on a stock Google Nexus™ 10 device

Video Processing APK

MULTICORE WARE

- MulticoreWare Transcoding/Processing Pipeline
  - Image filters implemented using RenderScript
  - Optimized for ARM + Mali-T600 GPU Compute



Filter	FPS (GPU+CPU vs CPU only)	Speed-up
Deshake (720p)	28 / 8	3.5x
Upscaling (720p to 1080p)	20 / 3	6.7x

Mali GPU Computing Ecosystem

COMPUTATIONAL PHOTOGRAPHY AND ADVANCED IMAGING

Morpho, apical, Aptina IMAGING, Synthesis Corporation, OmniVision, NOKIA

SERVICES, LIBRARIES AND TOOLS

MULTICORE WARE, ArcSoft, AcceleREyes, Amdahl software, PUX

HPC

MONT BLANC

COMPUTER VISION APPLICATIONS

crunchfish.com, EXTREME REALITY, metaio AUGMENTED SOLUTIONS, SoftKinetic, eyeSight™

MULTIMEDIA PROCESSING

Ittiam, DOLBY DIGITAL, SQUID DESIGN SYSTEMS, ARICENT GROUP, VisualOn, PIXTREE

OEM DIRECT ENGAGEMENTS

HUAWEI, amazon

Summary

- Compute more efficiently using heterogeneous and parallel processing
- Use OpenCL to enable portable heterogeneous multiprocessing
- Mali-T600 GPUs brings efficient GPU computing to you...now