ARM Mali GPU Architecture

Sam Martin
Graphics Architect, ARM

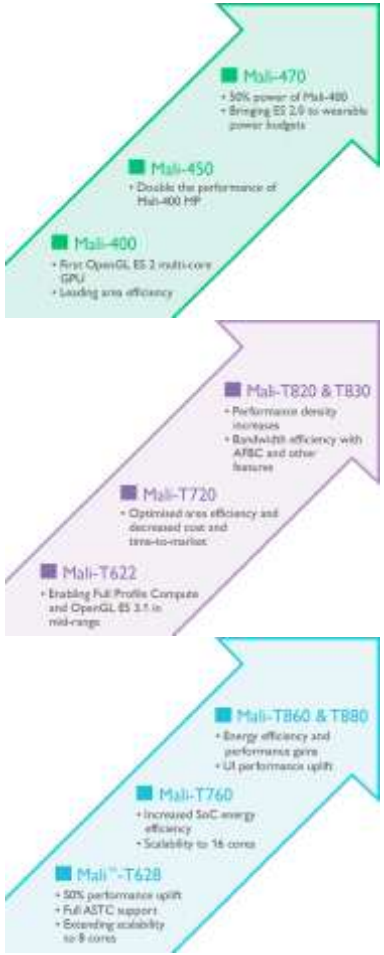
ARM Game Developer Day - London
03/12/2015

Agenda

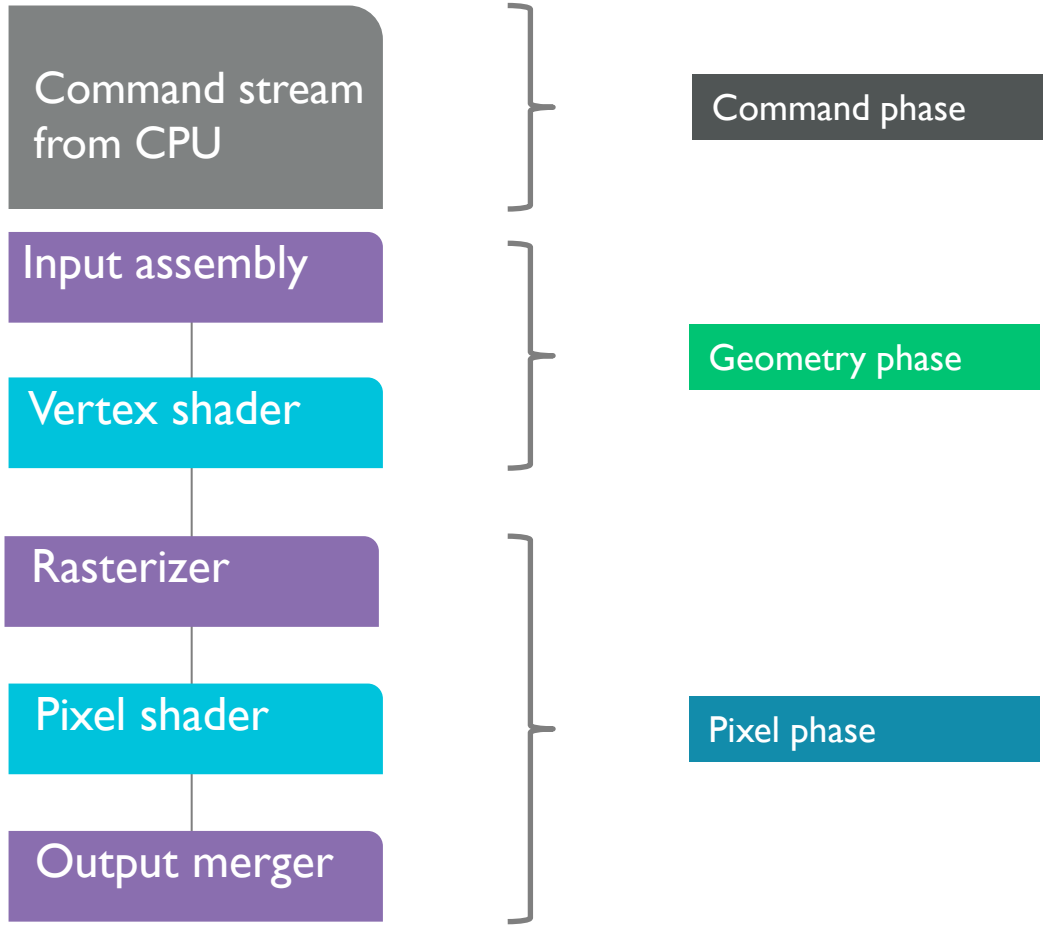
- Mali architecture and tiling introduction
- Behind the scenes – power limits
- Vulkan

Mali GPU Taxonomy In a Nutshell

- Mali 4xx series → OpenGL ES 2.0
  - 1-8 shaders cores, separate fragment and vertex processors
- Mali 6xx – 8xx → OpenGL ES 3.x
  - Unified "tri-pipe" shader core
  - Larger core configurations, max 16 cores from Mali 760 +
  - AFBC, ASTC, Transaction Elimination, ...
- All tile-based GPUs



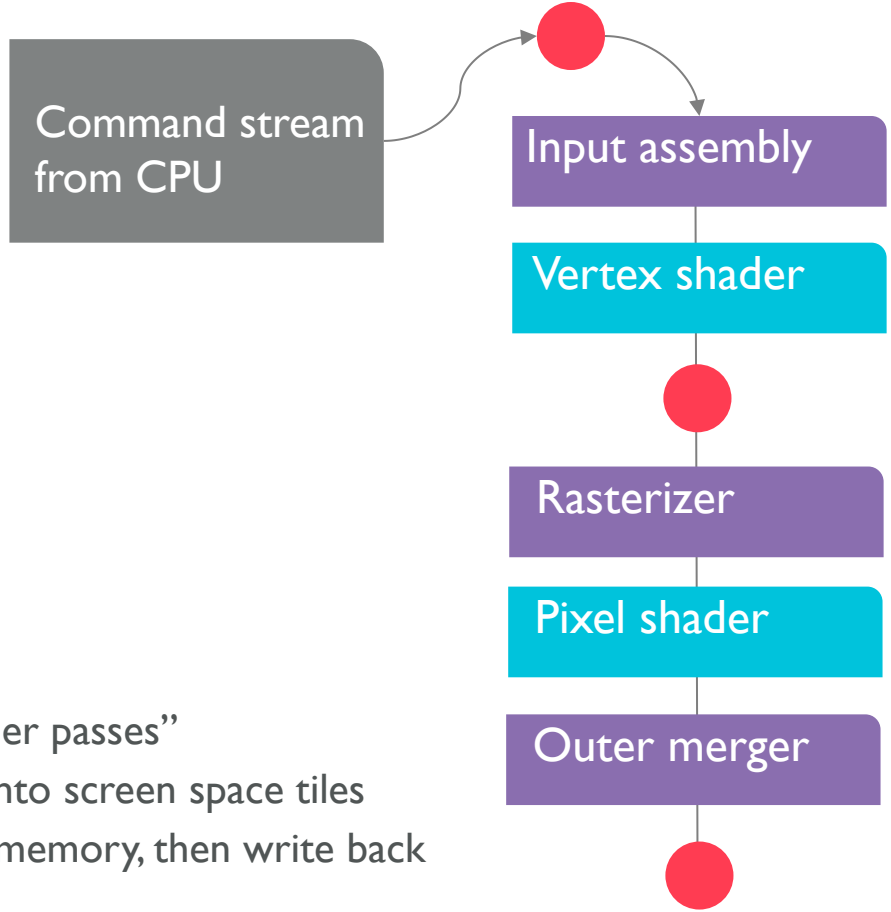
ARM


Command stream from CPU
Input assembly
Vertex shader
Rasterizer
Pixel shader
Output merger
Command phase
Geometry phase
Pixel phase

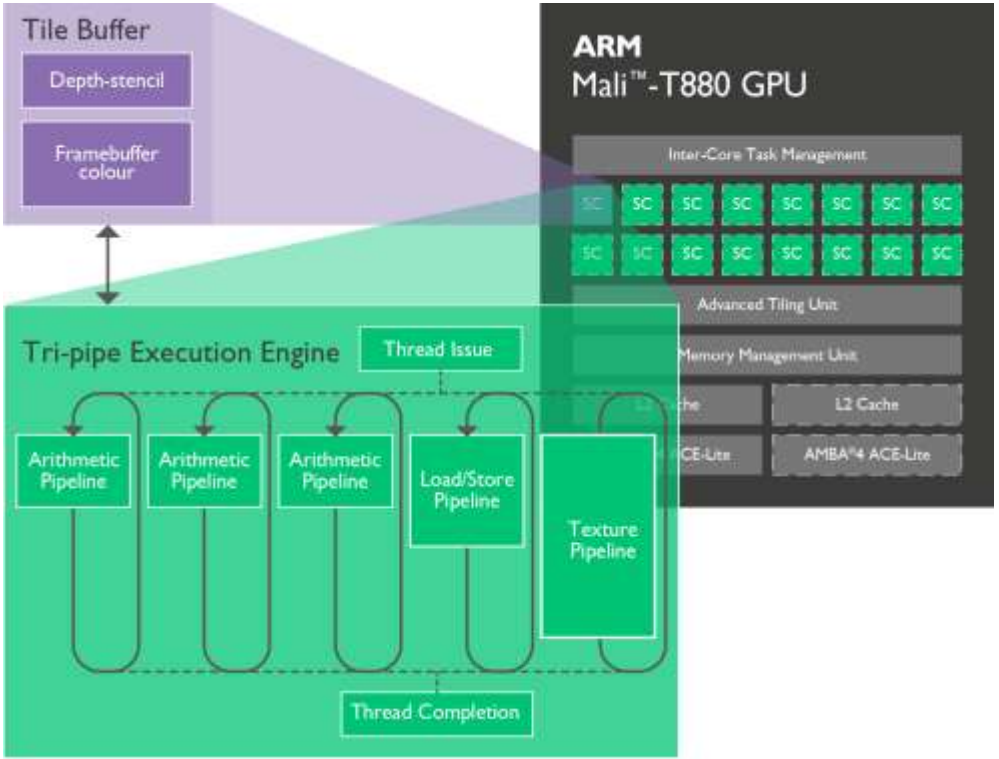
Tile-based GPUs

- Fragments >> Geometry

- Phased structure
  1. Buffer all operations into “render passes”
  2. Transform + bin all geometry into screen space tiles
  3. Fully shade each tile into local memory, then write back



Mali Architecture



- Hardware tiling
- Forward Pixel Kill
  - Reduce overdraw
- Framebuffer memory on-chip
  - 4x MSAA for “free”
  - Advanced on-chip shading
- Bandwidth efficiencies
  - ARM Framebuffer Compression
  - Transaction elimination
  - ASTC

ARM

Mobile Power Limits

- Lifetime constrained by battery
- High-end performance constrained by heat

- Thermal Design Power/Point (TDP)
  - Capacity constrained by ability to dissipate heat

- Memory bandwidth particularly expensive
  - Rule of thumb: 100mW / GB/s, assume 1 W total

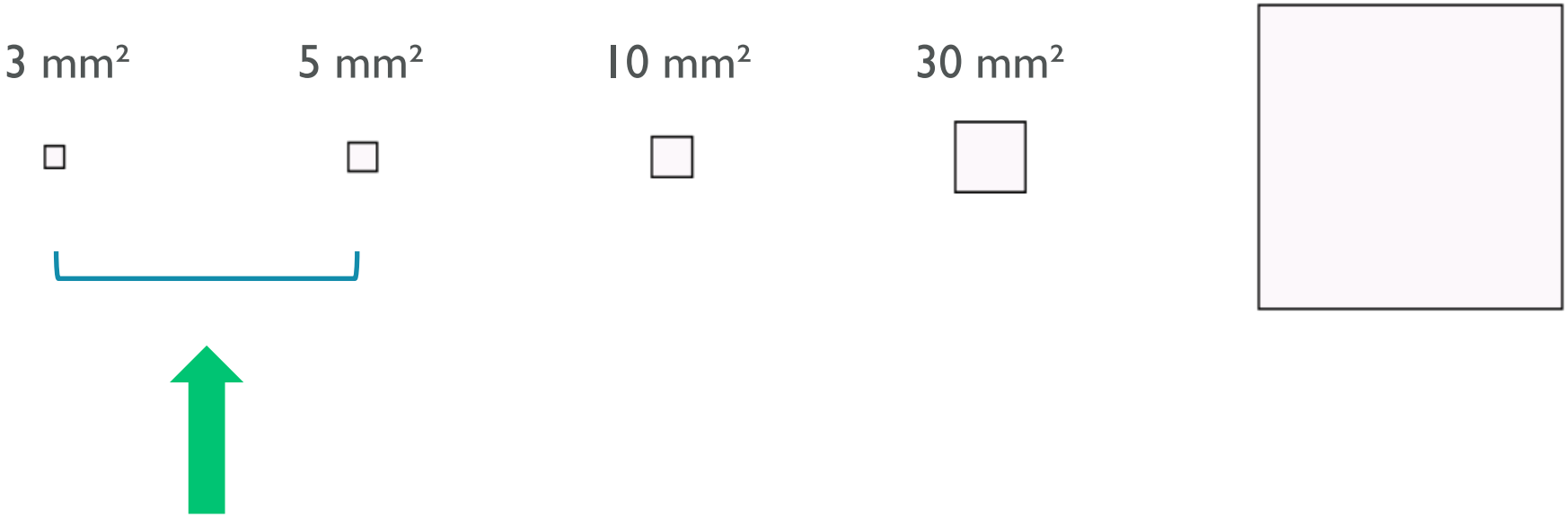
Phones	1-3 Watts
Tablets	3-5 Watts
Small laptop-like	10-25 Watts
Regular laptop	25-50 Watts
Integrated desktop	40-100 Watts

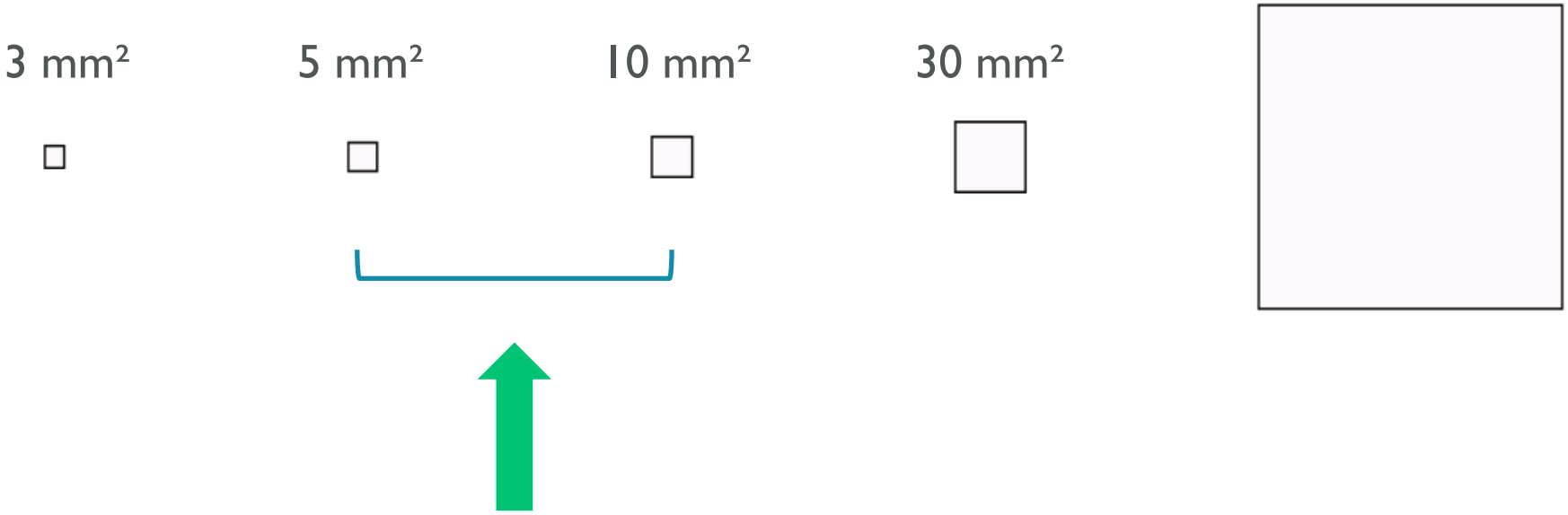
- Low-mid end GPUs are constrained by die area
  - Savings prolong battery life but may not increase performance

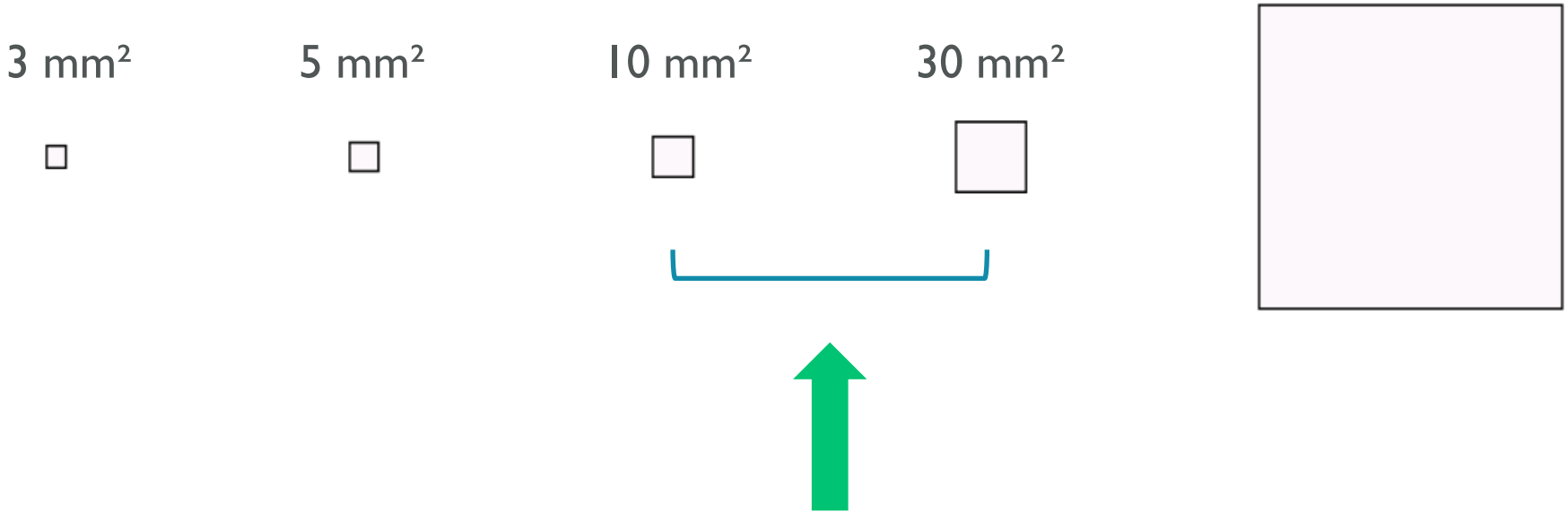
Similarly capable mobile GPUs
Die areas shown to scale

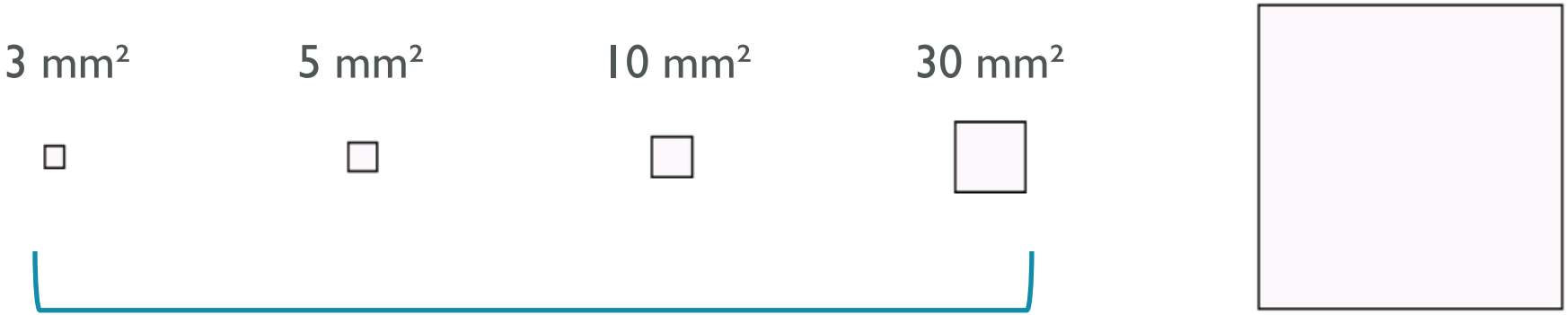
NVIDIA GeForce
GTX Titan

ARM







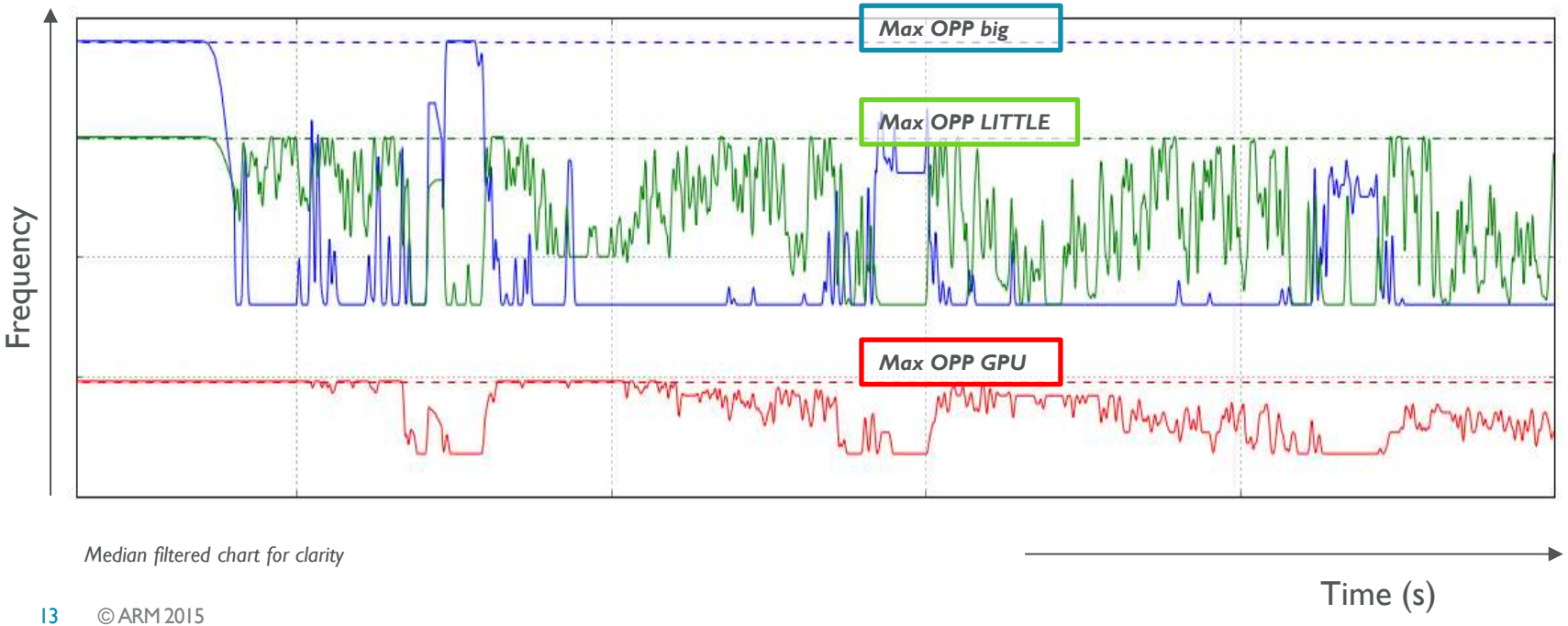


- 1-10x range, just within mobile phones
- Servicing such a wide range demands scalable GPU designs
- GPU feature set cannot indicate performance capability

Thermal Throttling

CPU - big
CPU - LITTLE
GPU

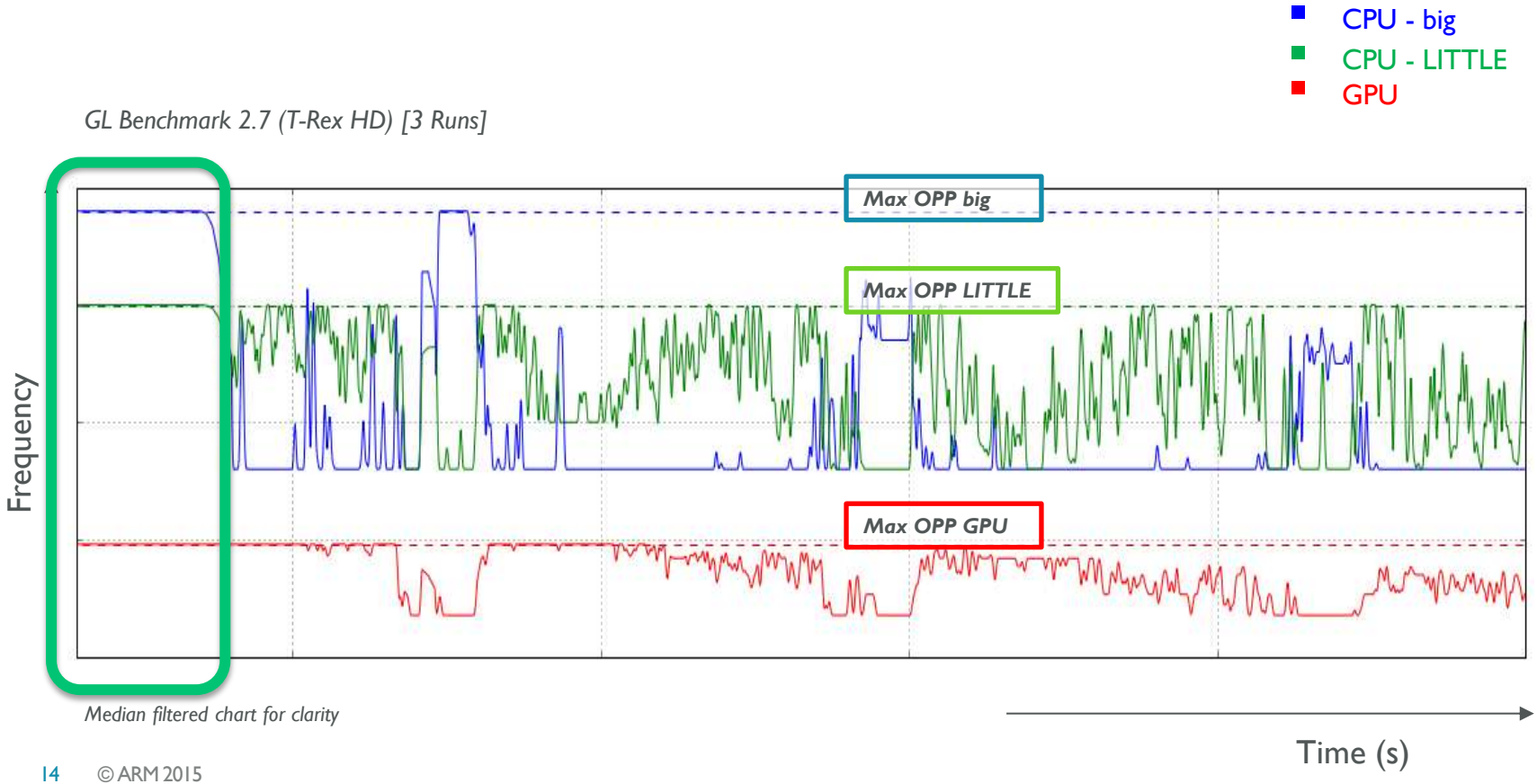
GL Benchmark 2.7 (T-Rex HD) [3 Runs]



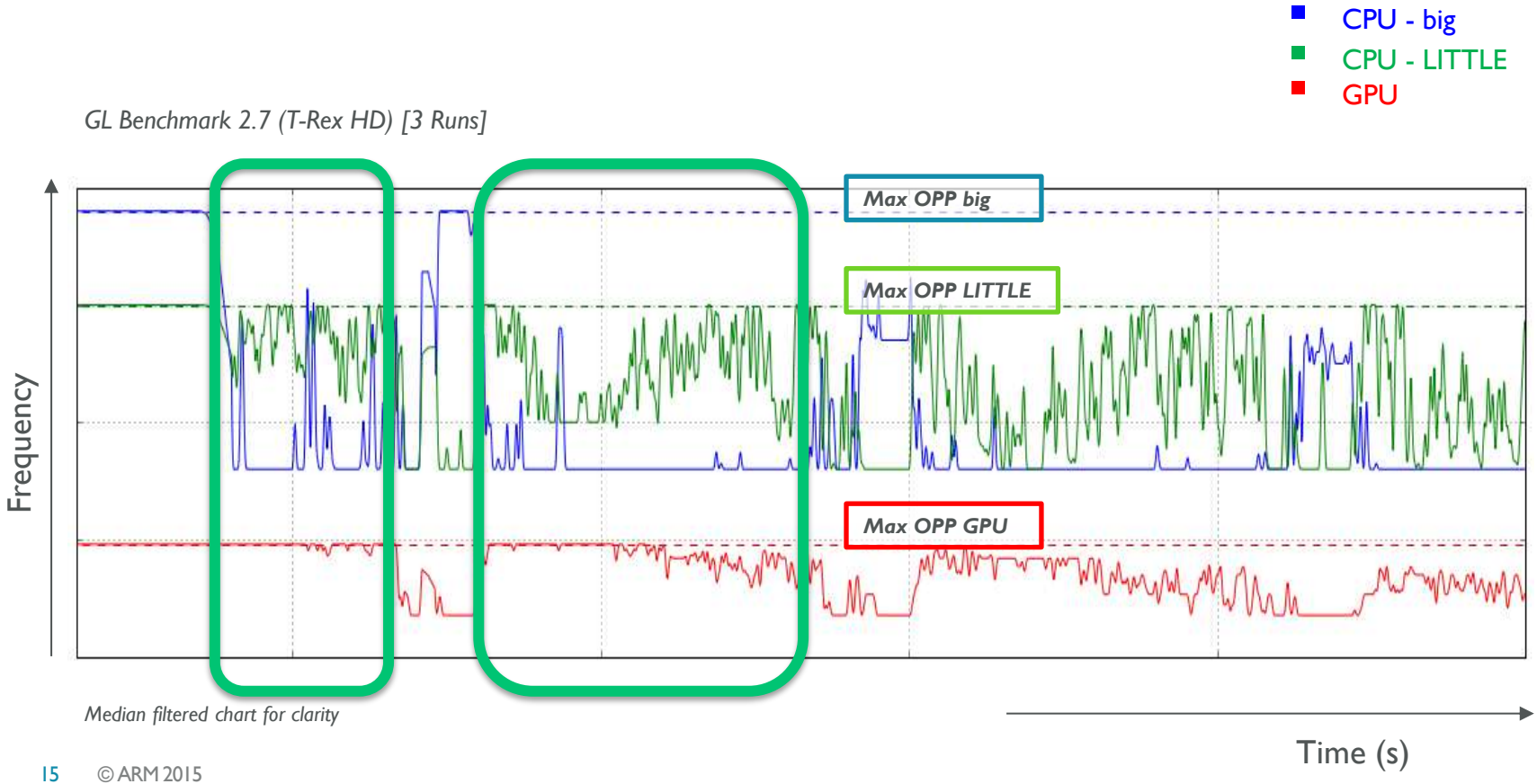
Median filtered chart for clarity

ARM

Thermal Throttling

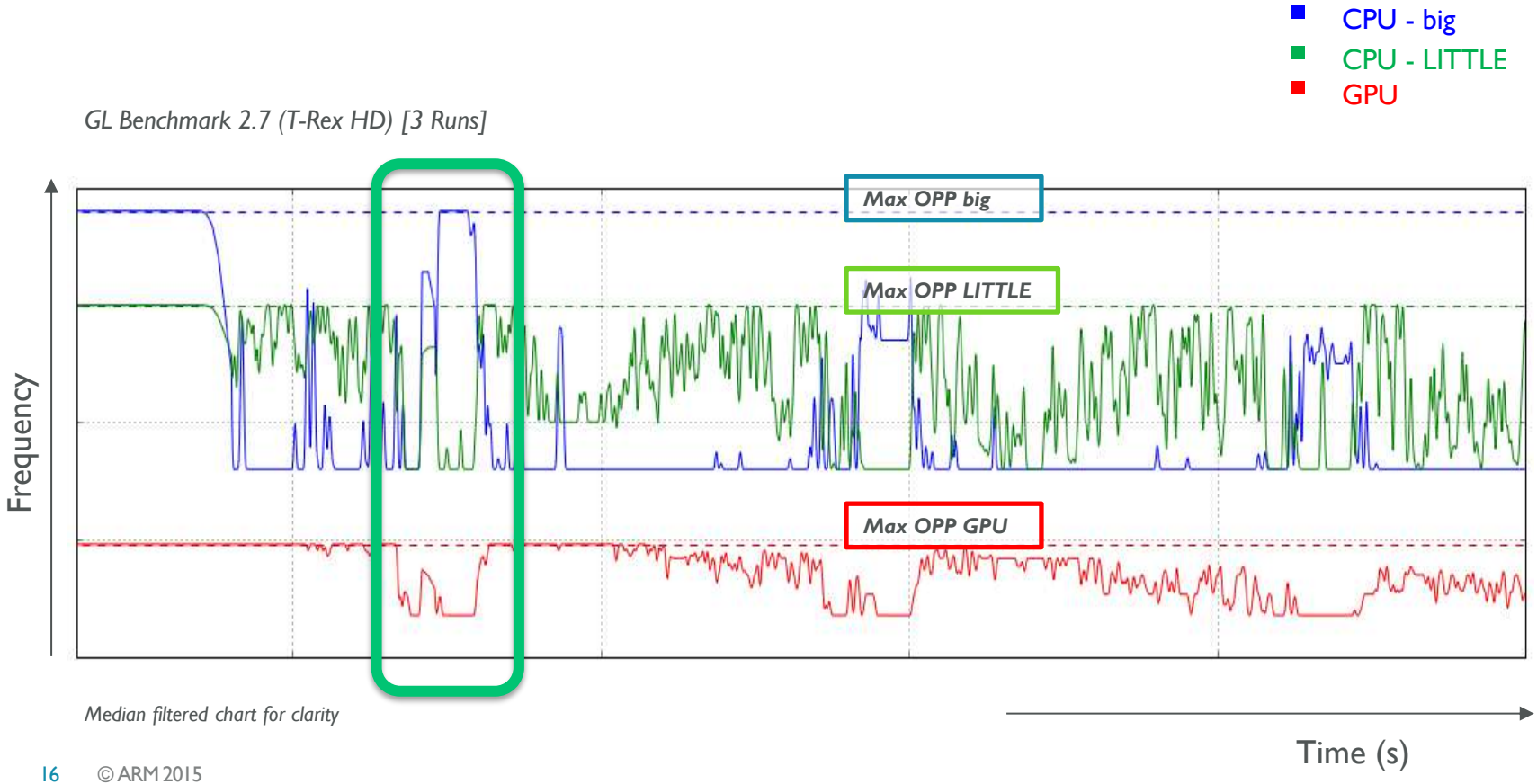


Thermal Throttling



ARM

Thermal Throttling

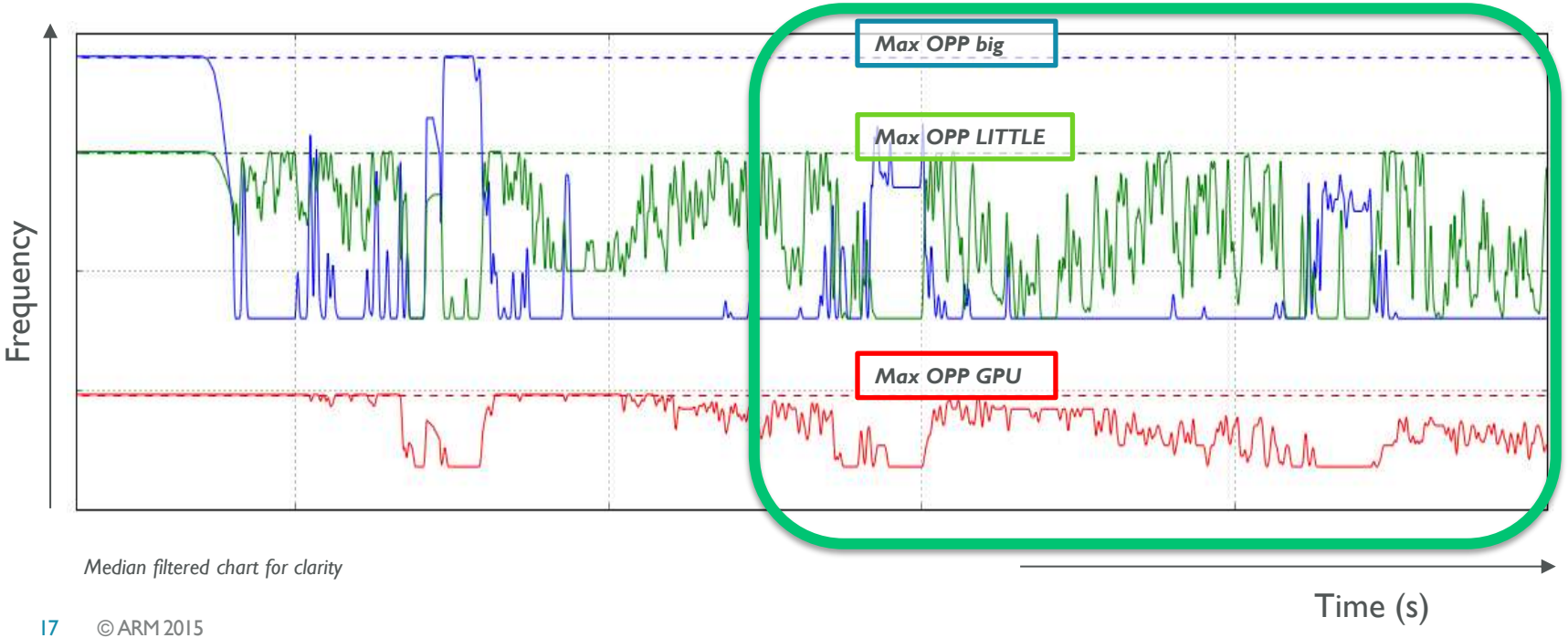


ARM

Thermal Throttling

CPU - big
CPU - LITTLE
GPU

GL Benchmark 2.7 (T-Rex HD) [3 Runs]



Median filtered chart for clarity

ARM

Vulkan

- Good match for mobile and tiling architectures
  - Explicit multi-pass render passes
  - No hidden costs (copies, allocs, shader recompiles, etc)
  - Multi-threaded
  - Low overhead

- Gloves-off API
  - Needs care – look out for future info post-release

Thanks! Questions?

sam.martin@arm.com @palgorithm

- Coming up:
- Increase texturing efficiency and quality
  - Daniele Di Donato, "Get the most out of ASTC" – up next!
- Advanced use of tiled framebuffers
  - Marius Bjørge, "Fast Approximate Indirect Lighting on Mobile", 11am
- Compute shaders & tessellation
  - Hans-Kristian Arntzen, "Real-time GPU-driven Ocean Rendering on Mobile", 11.30am

For more information visit the Mali Developer Centre:

http://malideveloper.arm.com

- Revisit this talk in PDF and audio format post event
- Download tools and resources

ARM