Mali-400 MP: A Scalable GPU for Mobile Devices

Tom Olson

Director, Graphics Research, ARM

Outline

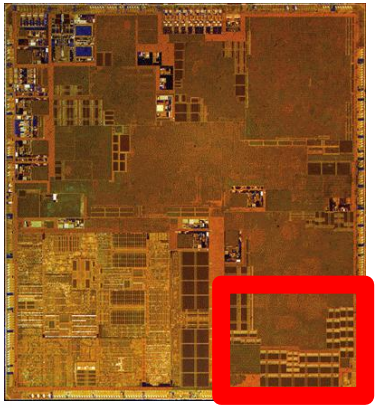
- ARM and Mobile Graphics
- Design Constraints for Mobile GPUs
- Mali Architecture Overview
- Multicore Scaling in Mali-400 MP
- Results

About ARM

- World's leading supplier of semiconductor IP
  - Processor Architectures and Implementations
  - Related IP: buses, caches, debug & trace, physical IP
  - Software tools and infrastructure
- Business Model
  - License fees
  - Per-chip royalties
- Graphics at ARM
  - Acquired Falanx in 2006
  - ARM Mali is now the world's most widely licensed GPU family

Challenges for Mobile GPUs

- Size

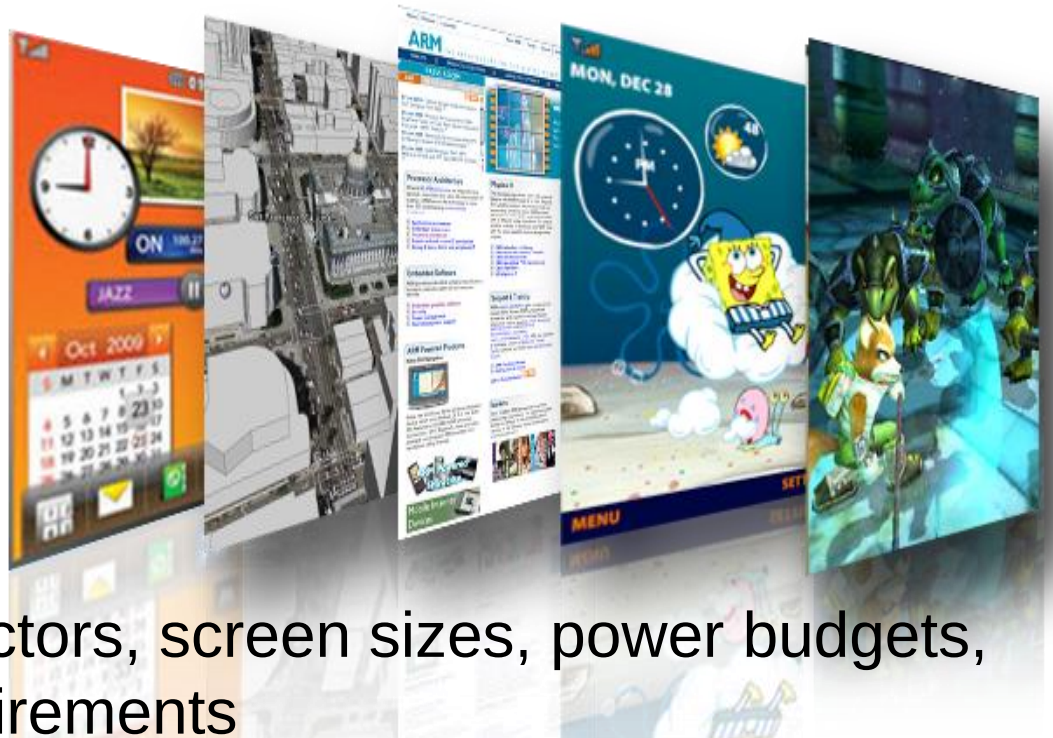


- Power

- Memory Bandwidth

More Challenges

- Graphics is going into “anything that has a screen”
  - Mobile
  - Navigation
  - Set Top Box/DTV
  - Automotive
  - Video telephony
  - Cameras
  - Printers



- Huge range of form factors, screen sizes, power budgets, and performance requirements
- In some applications, a huge difference between *peak* and *average* performance requirements

Solution: Scalability

- Address a wide variety of performance points and applications with a **single IP** and a **single software stack**.

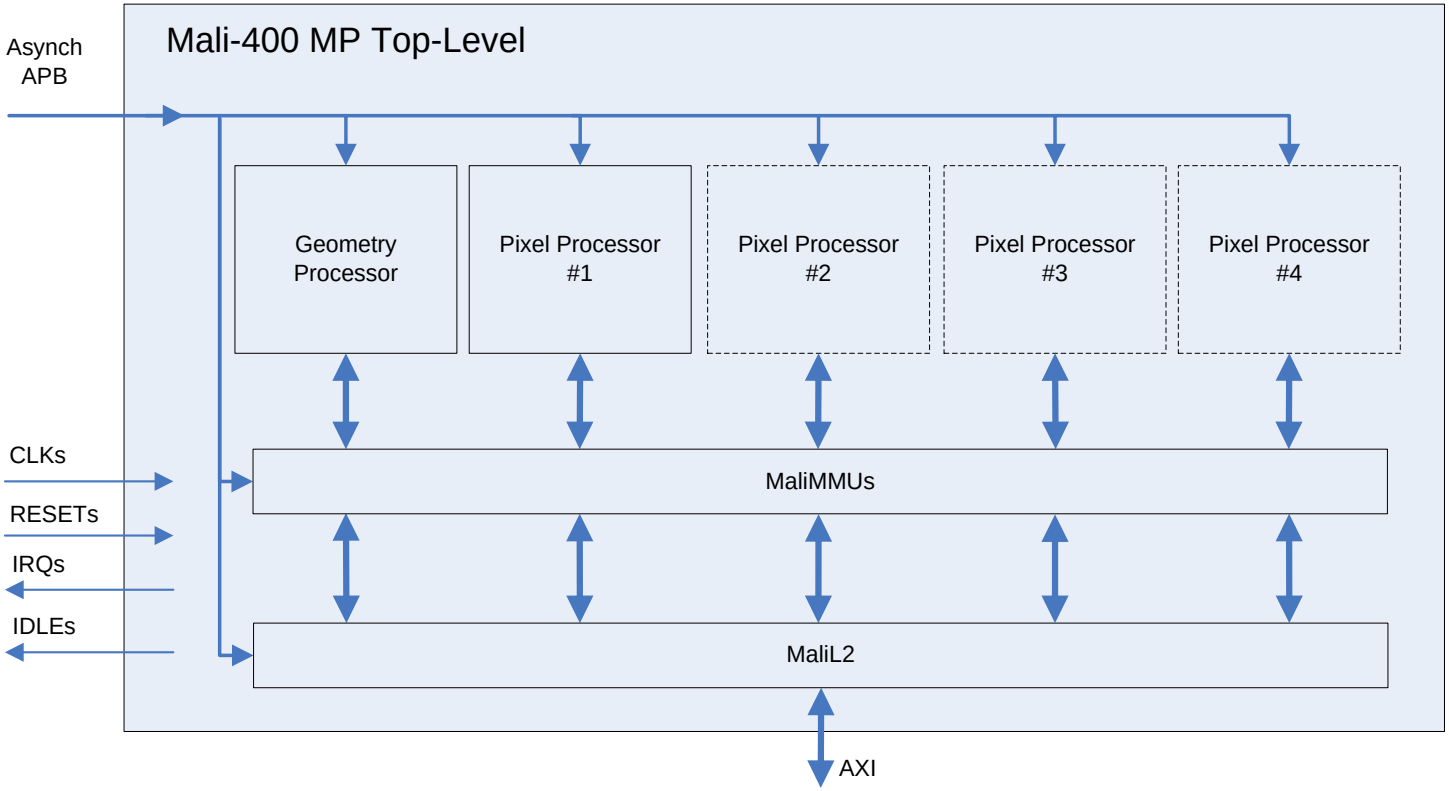
- Need *static* scalability to adapt to different peak requirements in different platforms / markets
- Need *dynamic* scalability to reduce power when peak performance isn't needed

Options for Scalability

- Fine-grained: Multiple pipes, wide SIMD, etc
  - Proven approach, efficient and effective
  - But, adding pipes / lanes is invasive
    - Hard for IP licensees to do on their own
  - And, hard to partition to provide dynamic scalability

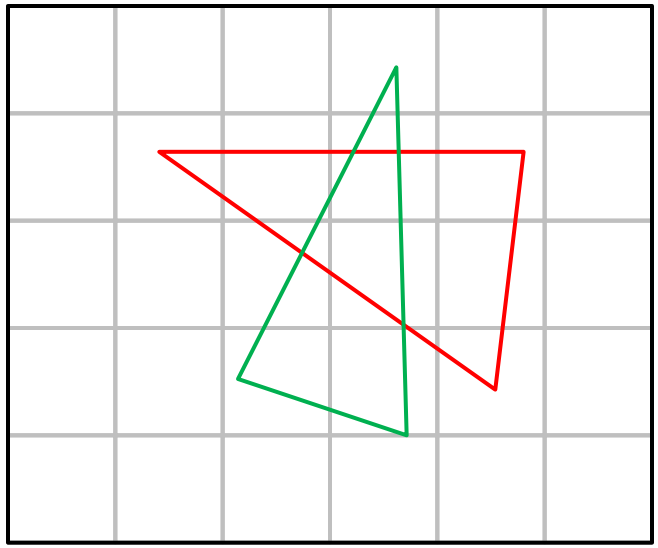
- Coarse-grained: Multicore
  - Easy for licensees to select desired performance
  - Putting cores on separate power islands allows dynamic scaling

Mali 400-MP Top Level Architecture



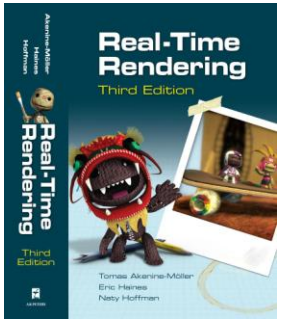
- Scalable pixel performance
  - 1-4 rasterizer cores
  - 32K-128K L2 cache

Mali Tile-Based Rendering

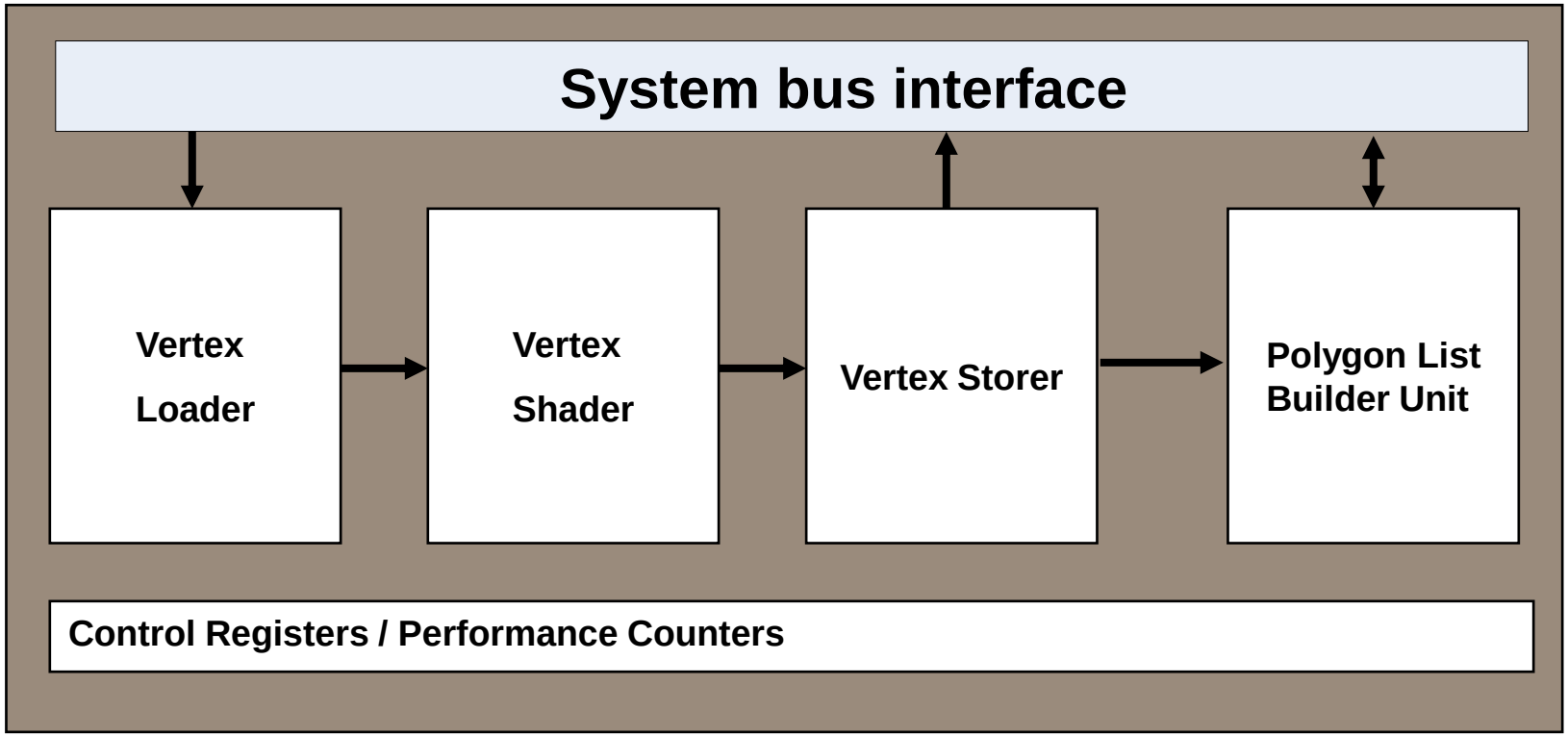


		1	1		
	0	0,1	0,1	0	
	0	0,1	0,1	0	
	0	0,1	0,1	0	

- Reduces off-chip framebuffer bandwidth
  - Rasterize into on-chip 16x16 tile buffer
  - Z, stencil, MSAA samples never go off-chip
  - Tradeoff against increased geometry bandwidth
- Details: see *Real-Time Rendering, 3rd ed.*



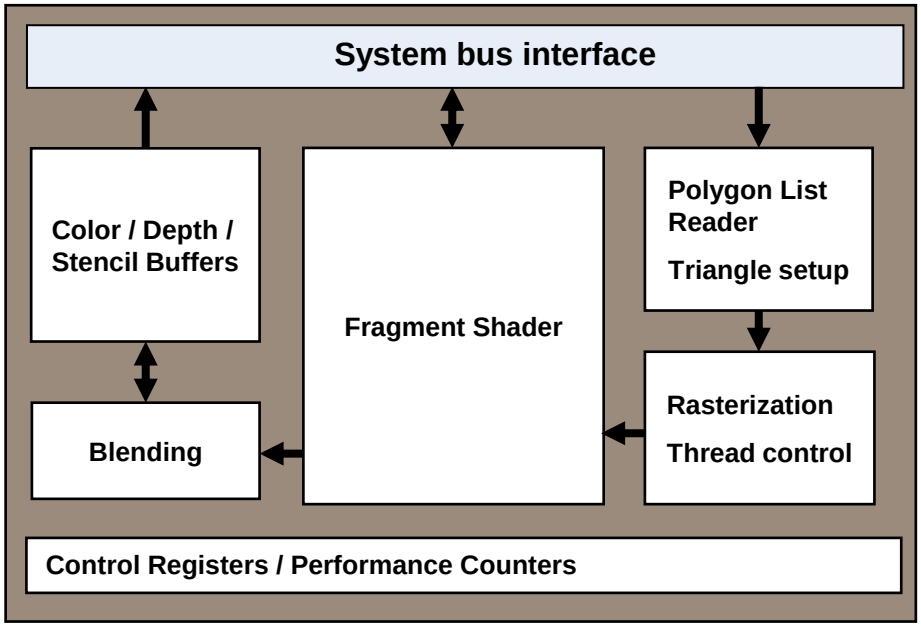
Mali-400 MP Geometry Processor



- Vertex Shader
  - Single-threaded, deeply pipelined

Mali-400 MP Pixel Processor

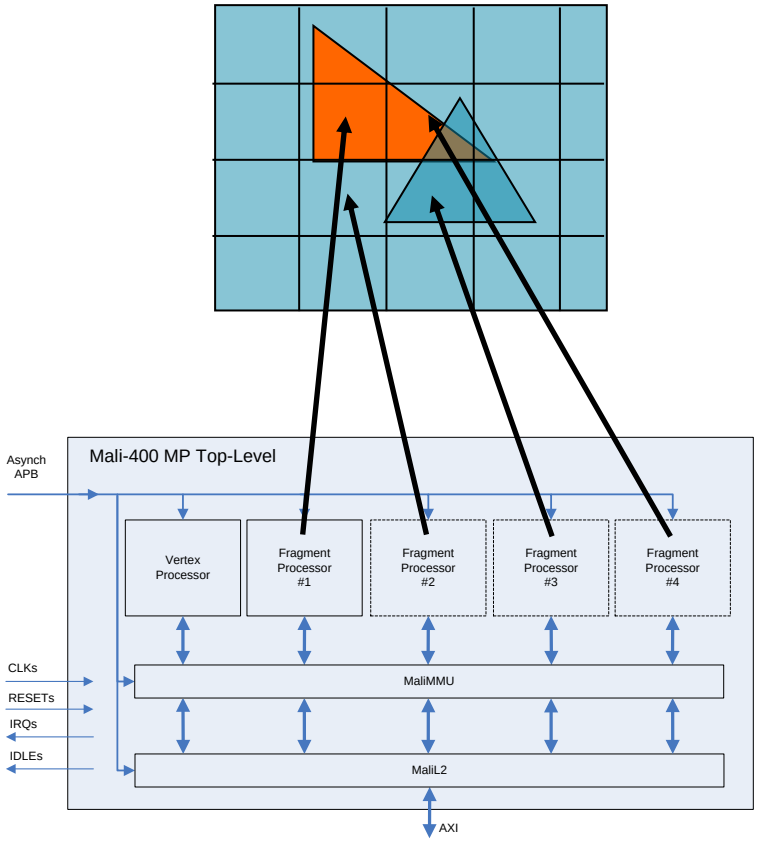
- Fragment Shader
  - 128-thread barrel processor
  - Fully general control flow
  - VLIW ISA, tuned for graphics
  - One texture sample per clock



- Key Features
  - 16x16 on-chip tile buffer
  - Renders one pixel per clock: 275M pix/sec @ 275 MHz
  - No penalty for 4x MSAA
  - No penalty for blending
  - 4x or 16x MSAA resolve on output
  - OpenGL ES 2.0 states handled without state dependent shaders

Going Multi-Core

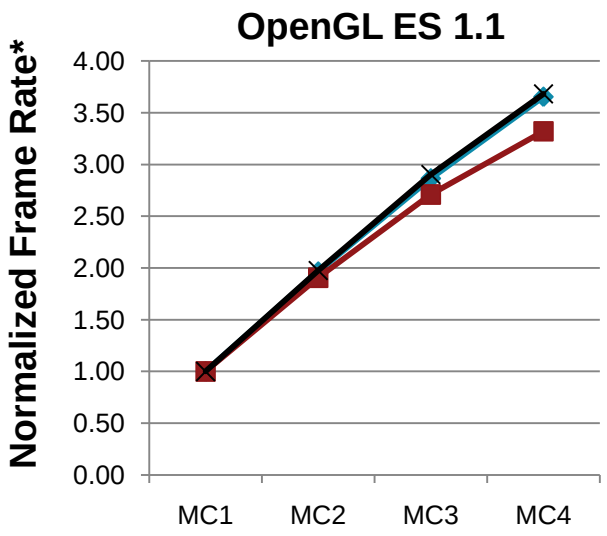
- All cores work in parallel on separate tasks
- Each core processes one tile at a time until completion – no communication between cores
- Tiles assigned statically to cores in a swizzled order
- Tile processing order maximizes L2 hit rate for polygon descriptors, textures

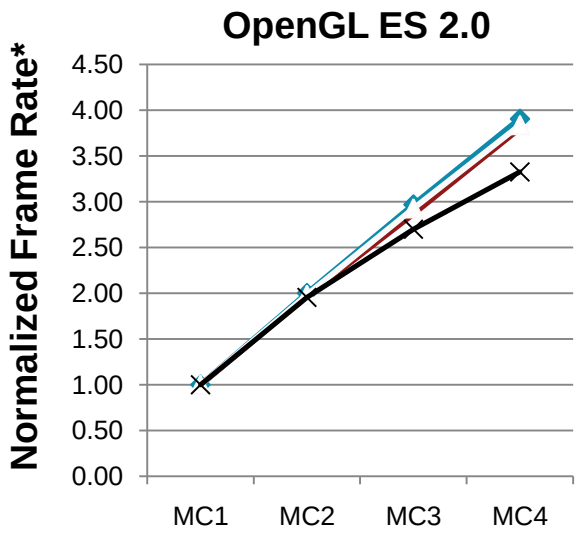


Does it work?

- Test content
  - Industry standard benchmarks for GL ES 1.1, 2.0
  - Thanks Rightware!

Relative Performance

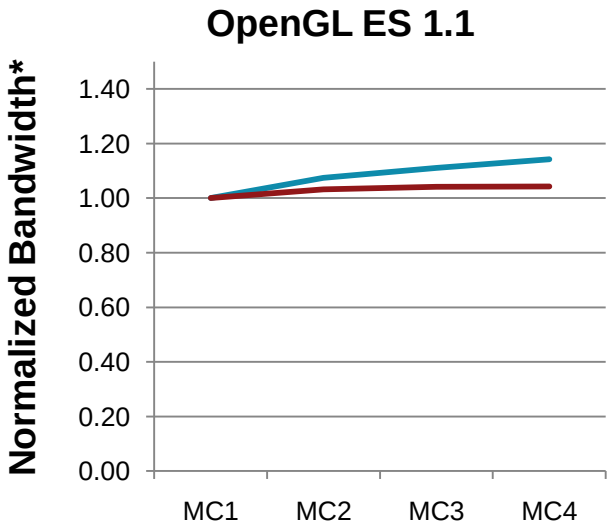




- Single frames in RTL simulation
- Artificial memory model (fixed static latency)
- 1-4 cores, 32KB L2 cache
- 1920x1080, 32bpp, 4xMSAA

***Single core frame rate = 1.0**

Impact on Bandwidth



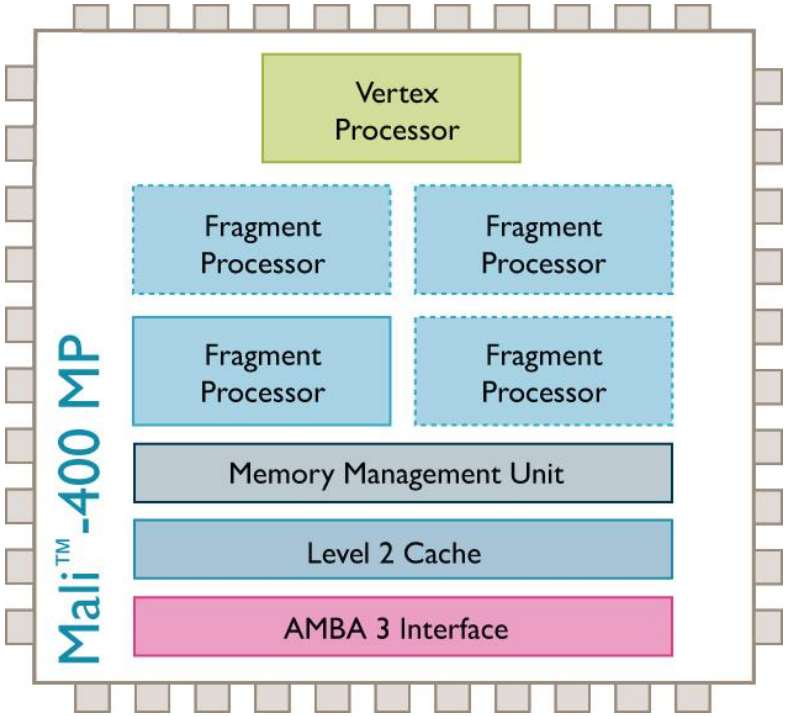
- Memory bandwidth per frame is stable
  - L2 cache is working to prevent redundant memory accesses

***Single core bandwidth per frame = 1.0**

Summary

- Scalability is an essential feature for mobile GPUs
  - Static *and* Dynamic
- Multi-core architectures are a good fit to the need
  - Easy to configure for different performance requirements
  - Easy to power-gate for dynamic scaling
- Mali-400 MP shows that the concept works
  - Pixel rate scalable from 275 Mpix/s to 1.1 Gpix/s
  - Linear speedup demonstrated on common mobile benchmarks
  - Single driver for all configurations – transparent to applications
  - Power and bandwidth efficient

Thanks to…



- The HPG 2010 Hot3D organizers
- Remi Pedersen, Mashhuda Glencross, and the Mali team
- Our Silicon Partners