Low power GPU computing The state of the union

Simon McIntosh-Smith simonm@cs.bris.ac.uk

Head of Microelectronics Research

University of Bristol, UK

University of BRISTOL

PEGPUM, HiPEAC Jan 2015

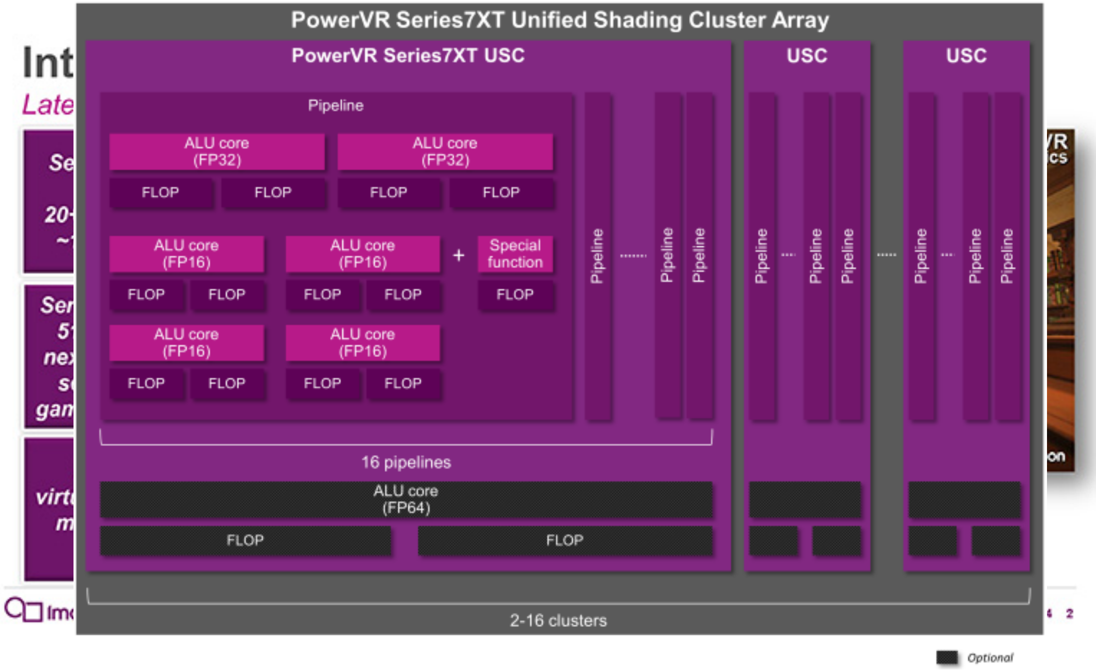
Considerable progress

Embedded, programmable, low-power GPUs have enjoyed a tremendous rate of progress in the last 12 months:

- Lots of new hardware and software products increasing in maturity
- Rapidly expanding ecosystems
- Increasing deployment into the mainstream

Progress in LPGPU hardware

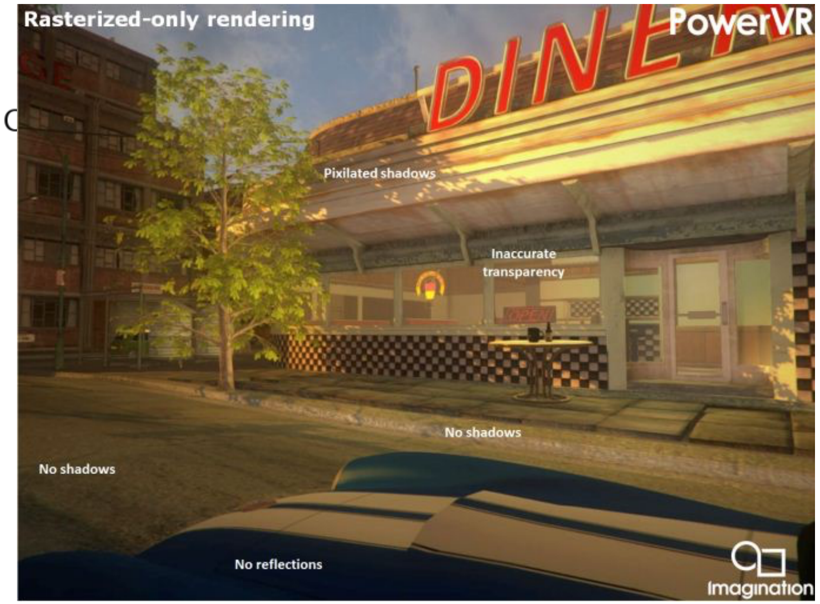
Imagination Technologies now has 64-bit MIPS cores and 64-bit floating point capable PowerVR GPUs up to 1.5 TFLOP/s (single precision) [1,2]

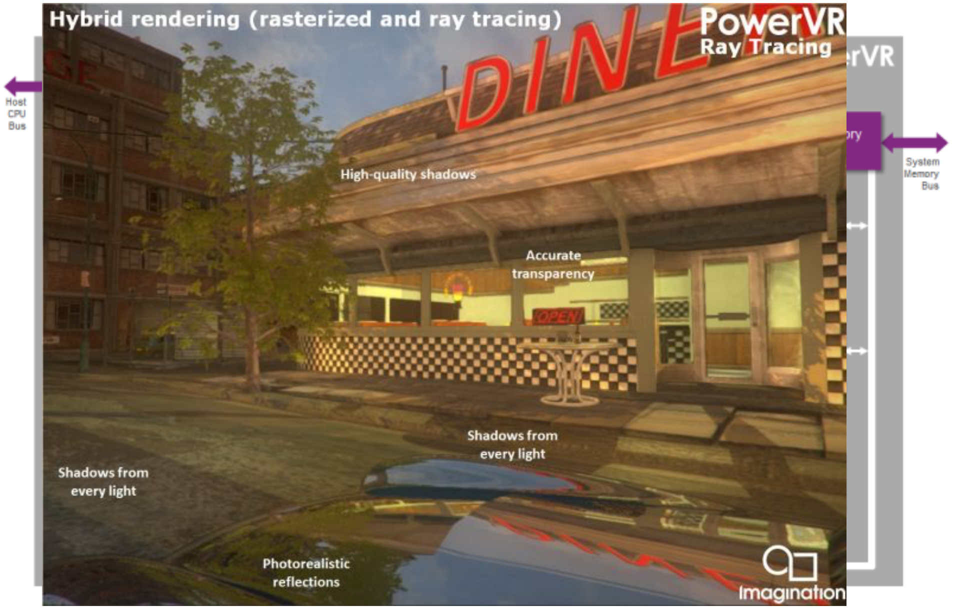


Sources:
[1] http://www.anandtech.com/show/8457/mips-strikes-back-64bit-warrior-i6400-architecture-arrives/4
[2] http://www.anandtech.com/show/8706/imagination-announces-powervr-series7-gpus-series7xt-series7xe

Progress in LPGPU hardware

Embedded GPUs extend to new application areas such as ray tracing [1] and video [2]

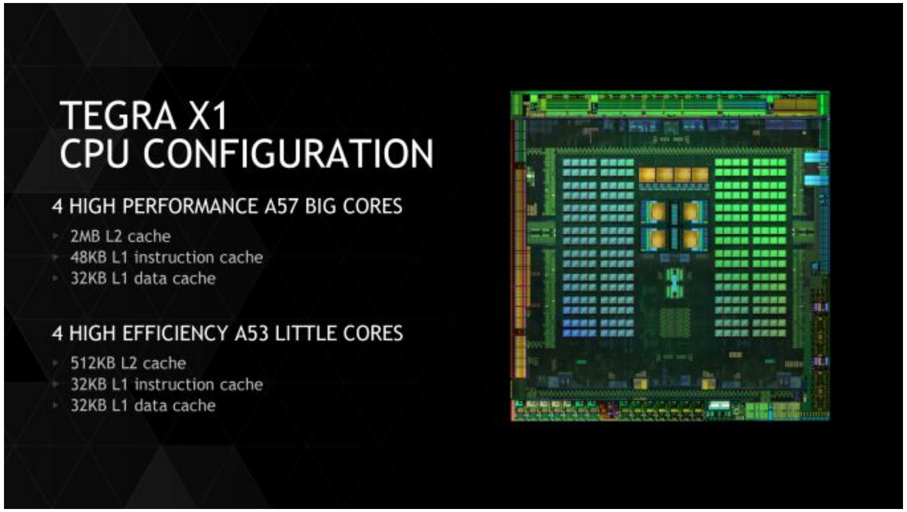




Sources:
[1] http://www.anandtech.com/show/7870/imagination-announces-powervr-wizard-gpu-family-rogue-learns-ray-tracing
[2] http://www.tomshardware.com/news/arm-mali-gpus-video-display,27961.html

Progress in LPGPU hardware

Embedded GPUs becoming compute monsters; Nvidia's Tegra X1 integrates eight 64-bit ARM cores (4+4 in big.LITTLE configuration), a GPU capable of 0.5 TFLOP/s single precision and up to 25.6 GBytes/s of memory bandwidth for ~10W [1]



TEGRA X1
CPU CONFIGURATION

4 HIGH PERFORMANCE A57 BIG CORES

- 2MB L2 cache
- 48KB L1 instruction cache
- 32KB L1 data cache

4 HIGH EFFICIENCY A53 LITTLE CORES

- 512KB L2 cache
- 32KB L1 instruction cache
- 32KB L1 data cache

Sources:
[1] http://www.anandtech.com/show/8811/nvidia-tegra-x1-preview

Progress in software

Significant progress in software for embedded GPUs:

- **Standards** such as OpenCL 2.0 and SPIR
- Increasing software **ecosystem**
- LPGPU **applications** emerging (HDR etc.)
- Continued innovation in **programming languages** – Apple's Metal / OpenGL-next

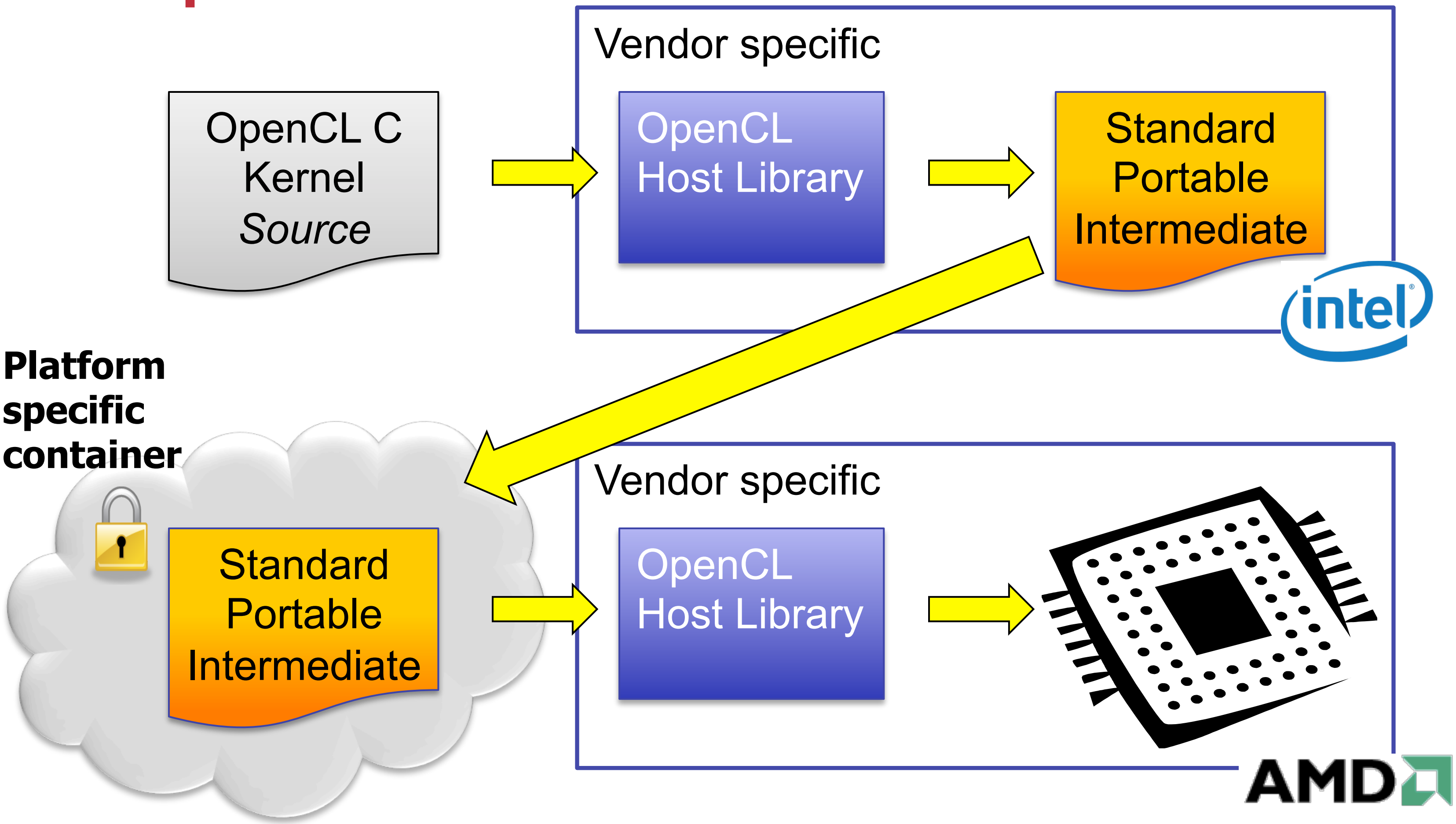
Progress in software

OpenCL 2.0 & SPIR enable new tools for developers:

- **OpenCL 2.0** adds significant new features that benefit embedded GPUs [1]:
  - Shared virtual memory (SVM) between CPU and GPU
  - Nested parallelism (device can enqueue kernels)
  - Built-in functions for reductions, broadcasts …
- **SPIR 1.2** radically lowers the barrier to entry to the OpenCL ecosystem [2]:
  - Can generate SPIR on one platform and use it on another
  - Can ship applications in portable binary format rather than as human readable kernel source code

Sources:
[1] https://www.khronos.org/opencl/
[2] https://www.khronos.org/spir

OpenCL: SPIR flow



- ISV ships kernels in SPIR form
- User runs application on platform of their choice

University of BRISTOL

Progress in software

OpenCL SPIR enables the creation of new tools for developers:

E.g. **Oclgrind**, an OpenCL device simulator

- Developed at the University of Bristol
- https://github.com/jrprice/Oclgrind/wiki

Oclgrind

- Simulates OpenCL kernels executing on a virtual OpenCL device
- Built on an interpreter for SPIR
- Architecture-agnostic simulation
- Plugin interface for ***extensibility***
- Has found bugs in substantial codes: Parboil, CloverLeaf, ViennaCL etc
- Extended by Codeplay to profile memory accesses

- http://www.many-core.group.cam.ac.uk/ukmac2014/UKMAC2014_07_Price.pdf

https://github.com/jrprice/Oclgrind/wiki

LPGPU applications emerging

As programmable GPUs start shipping in products, **applications** that use them are starting to appear, e.g.

- Computational photography pipelines (HDR etc)
- Image manipulation applications
- Deep learning / artificial neural networks
- Automotive
- Games!

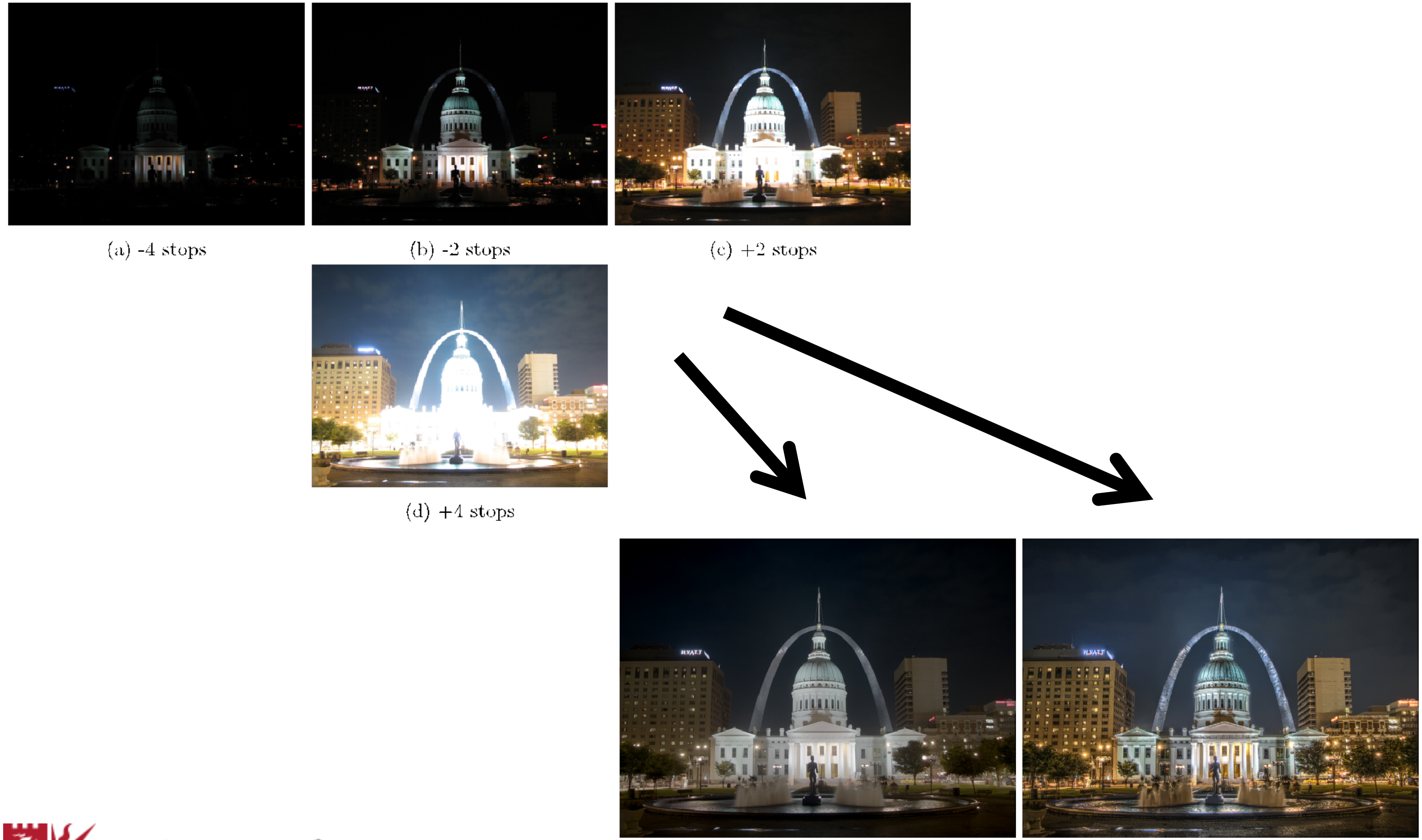
An HDR pipeline for LPGPUs

The University of Bristol has developed a **high dynamic range computational photography pipeline** for OpenCL devices:

- Combines multiple images with a local or global tone mapping operator to enhance detail in areas of the image at the extremes of the exposure
- Achieves 30fps for 1920x1080 images on an ARM Mali T604 GPU for a Reinhard Global TMO
- To appear in *GPU Pro 6: Advanced Rendering Techniques,* Wolfgang Engel (ed.), March 2015.

- https://github.com/amirchohan/HDR

HDR image processing



(a) -4 stops

(b) -2 stops

(c) +2 stops

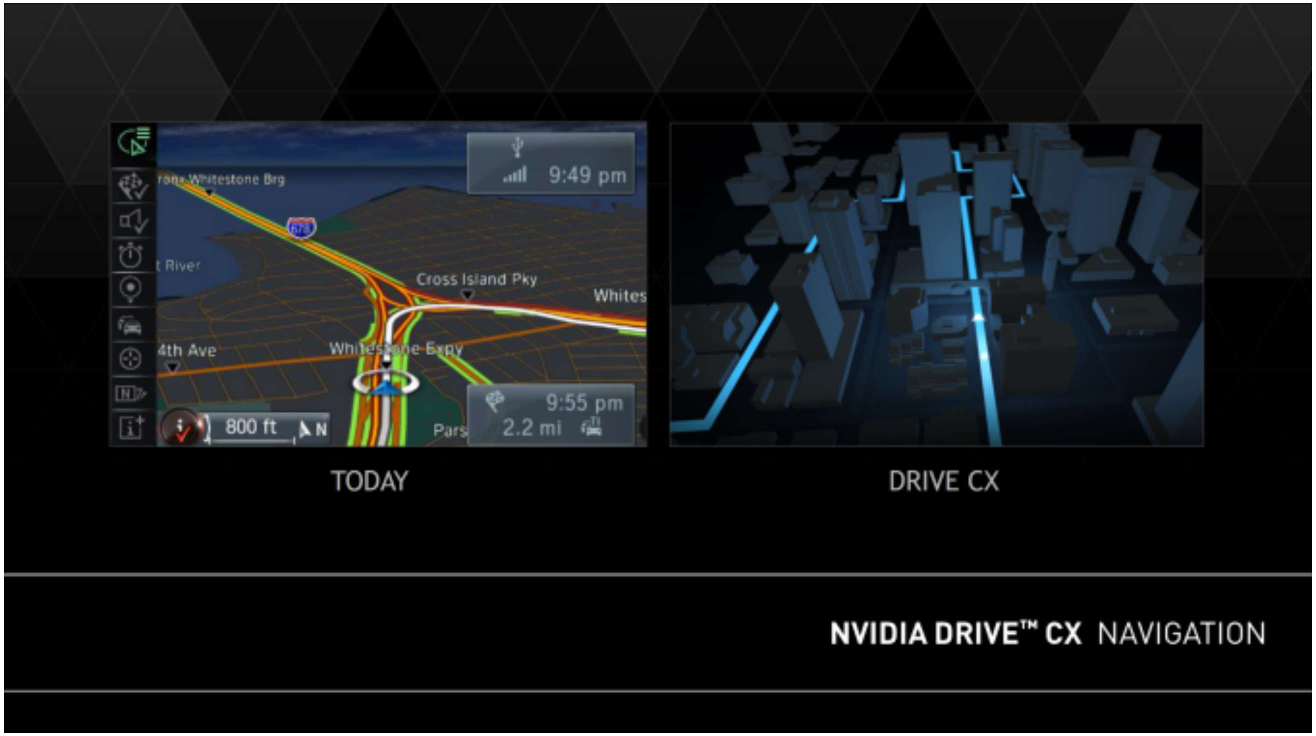
(d) +4 stops

(a) Global TMO

(b) Local TMO

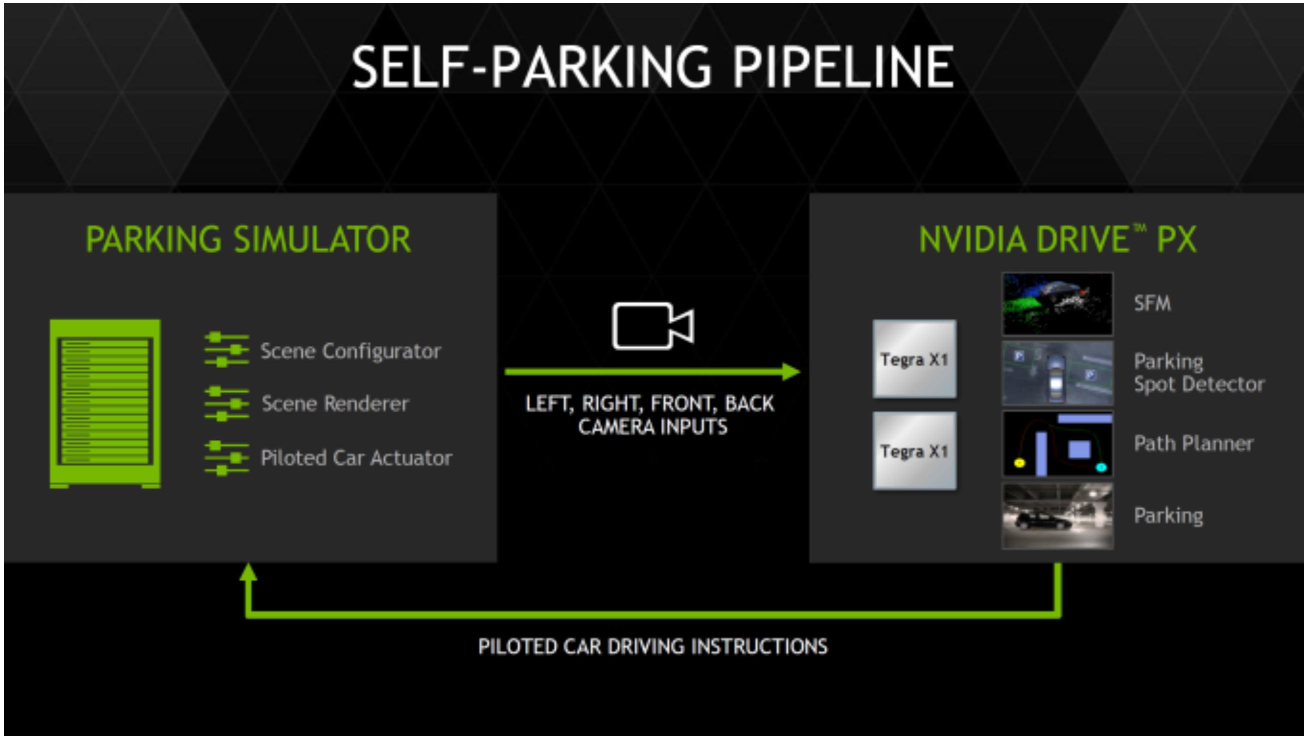
Automative / deep learning

Nvidia's making a lot of noise about LPGPUs for automotive and deep learning



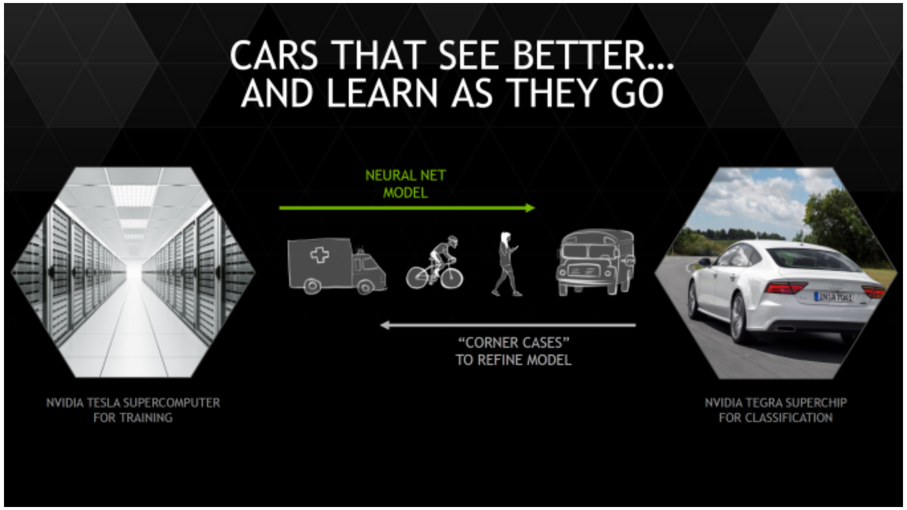
Automative / deep learning

Nvidia's making a lot of noise about LPGPUs for automotive and deep learning



Automative / deep learning

Nvidia's making a lot of noise about LPGPUs for automotive and deep learning



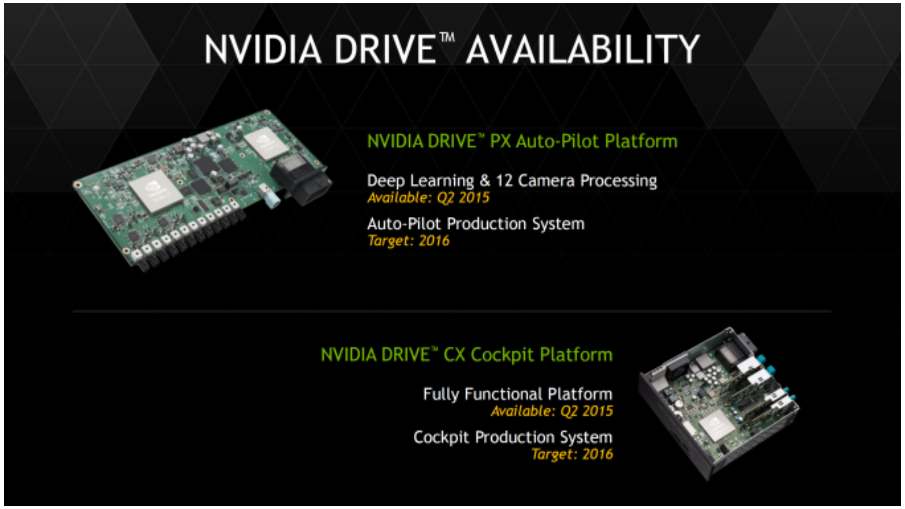
CARS THAT SEE BETTER...
AND LEARN AS THEY GO

NEURAL NET MODEL

"CORNER CASES" TO REFINE MODEL

NVIDIA TESLA SUPERCOMPUTER FOR TRAINING

NVIDIA TEGRA SUPERCHIP FOR CLASSIFICATION



NVIDIA DRIVE™ AVAILABILITY

NVIDIA DRIVE™ PX Auto-Pilot Platform

Deep Learning & 12 Camera Processing
Available: Q2 2015

Auto-Pilot Production System
Target: 2016

NVIDIA DRIVE™ CX Cockpit Platform

Fully Functional Platform
Available: Q2 2015

Cockpit Production System
Target: 2016

Progress in software: Metal

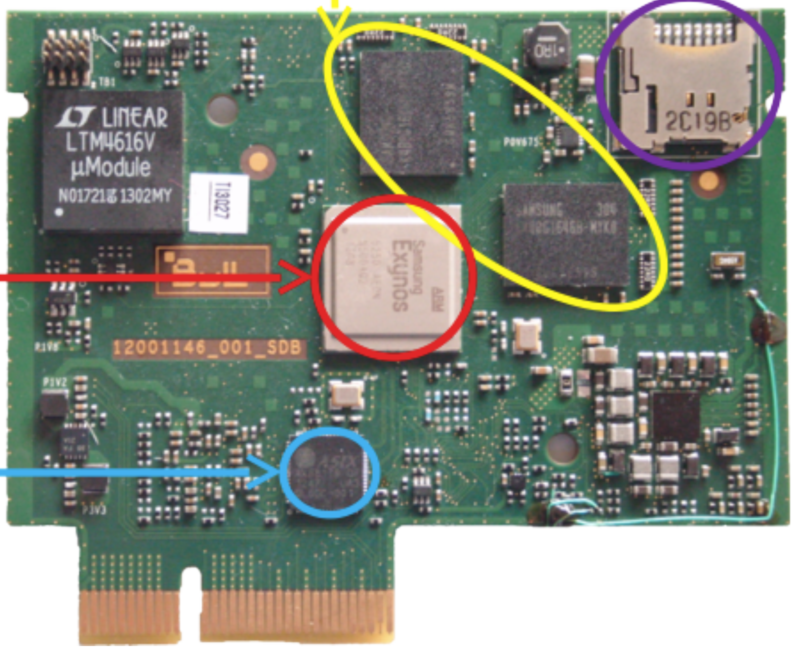
- Apple has just shaken up the GPU computing space by announcing their **Metal** API at WWDC in June 2014
- Defines a much lighter weight, higher performance graphics API than OpenGL
- Integrates compute capability with much lower switching overhead than OpenCL/ OpenGL interoperability
- Causing tremendous creative activity within the Khronos community!

For a recent talk I gave on Metal see: http://www.cs.bris.ac.uk/~simonm/publications/multicore_challenge_parallel_languages_Sep_2014.pdf

LPGPU for HPC – Mont Blanc

CPU + GPU + DRAM + storage + network
all in a compute card just 8.5 x 5.6 cm

MONT-BLANC



4 GB
DDR3-1600

µSDslot
up to 64 GB

Exynos5 Dual:
2x ARM Cortex-A15
ARM Mali-T604

USB 3.0
to 1GbE
bridge

University of
BRISTOL

http://www.montblanc-project.eu

Exynos 5 compute card

2 x Cortex-A15 @ 1.7GHz

1 x Mali T604 GPU

6.8 + 25.5 GFLOPS

15 Watts

2.1 GFLOPS/W

GPU ~ 2/3 peak
CPU ~ 1/3 peak

Carrier blade

15 x Compute cards

485 GFLOPS

1 GbE to 10 GbE

300 Watts

1.6 GFLOPS/W

Blade chassis 7U

9 x Carrier blade

135 x Compute cards

4.3 TFLOPS

2.7 kWatts

1.6 GFLOPS/W

Rack

6 BullX chassis

54 Compute blades

810 Compute cards

1620 CPU

810 GPU

3.2 TB of DRAM

52 TB of Flash

26 TFLOPS

18 kWatt

	Mont-Blanc [GFLOPS/W]	Green500 [GFLOPS/W]
Nov 2011	0.15	2.0
Jun 2014	1.5	4.4

Conclusions

- LPGPU is now fast maturing as a field
- Compelling hardware becoming available
- Software ecosystem becoming more vibrant
  - E.g. Oclgrind
- Applications that exploit the available hardware and software are emerging
  - E.g. HDR computational photography, automotive etc
- Increasing competition
  - E.g. Apple's Metal, Khronos glNext
- Lots of challenges remain - exciting times ahead!!

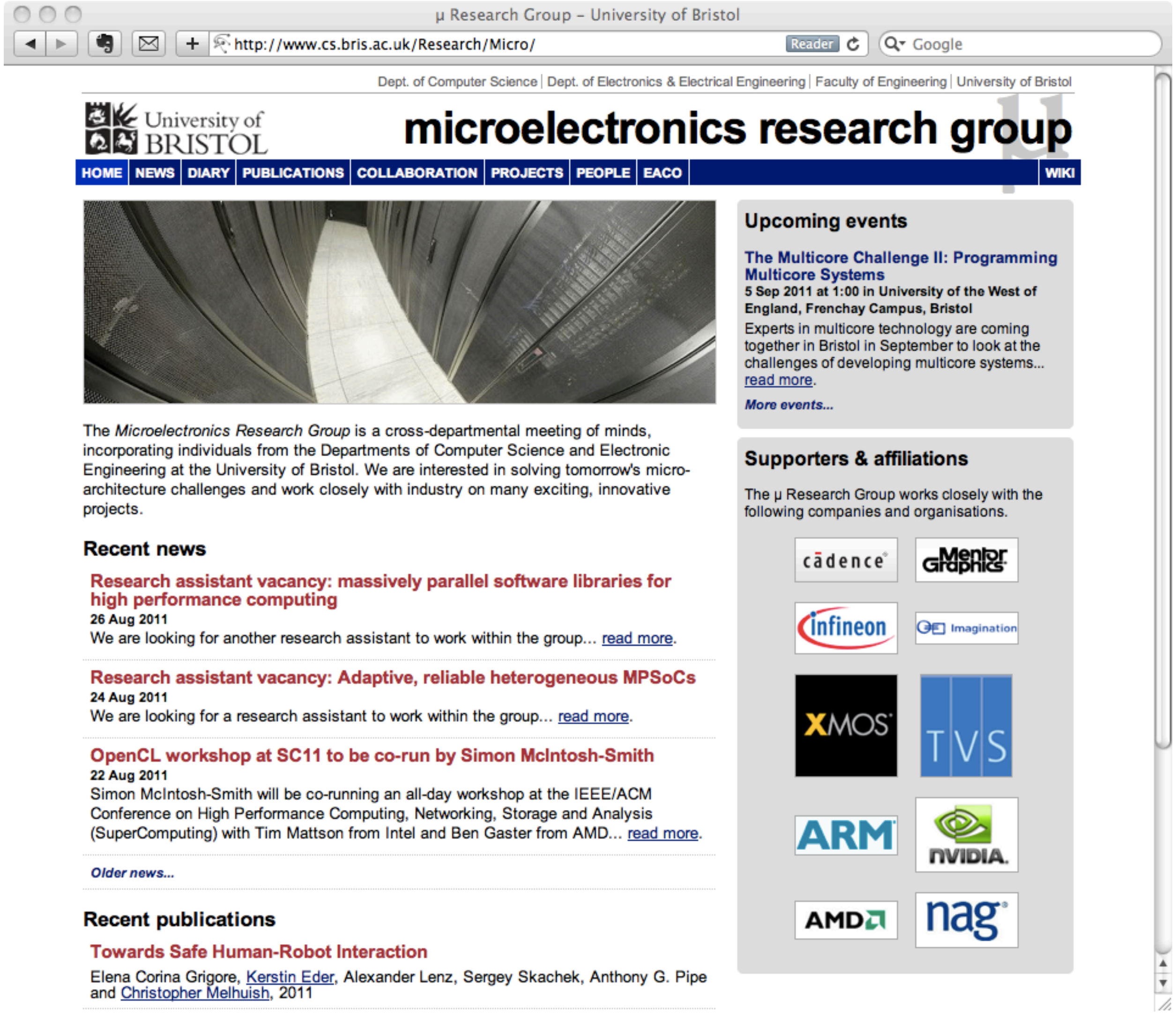
Shameless plug…

- The CFP for the 3rd International Workshop on OpenCL (IWOCL 2015) is open until February 14th
- This year at Stanford University, California May 12-13th

IWOCL
INTERNATIONAL WORKSHOP ON OPENCL

http://www.iwocl.org/

www.cs.bris.ac.uk/Research/Micro



μ Research Group – University of Bristol
http://www.cs.bris.ac.uk/Research/Micro/
Reader
Google
Dept. of Computer Science | Dept. of Electronics & Electrical Engineering | Faculty of Engineering | University of Bristol
University of BRISTOL
HOME
NEWS
DIARY
PUBLICATIONS
COLLABORATION
PROJECTS
PEOPLE
EACO
WIKI
cadence
Mentor Graphics
infineon
Imagination
XMOS
TVS
ARM
NVIDIA
AMD
nag

microelectronics research group

The *Microelectronics Research Group* is a cross-departmental meeting of minds, incorporating individuals from the Departments of Computer Science and Electronic Engineering at the University of Bristol. We are interested in solving tomorrow's micro-architecture challenges and work closely with industry on many exciting, innovative projects.

Recent news

Research assistant vacancy: massively parallel software libraries for high performance computing
26 Aug 2011
We are looking for another research assistant to work within the group... read more.

Research assistant vacancy: Adaptive, reliable heterogeneous MPSoCs
24 Aug 2011
We are looking for a research assistant to work within the group... read more.

OpenCL workshop at SC11 to be co-run by Simon McIntosh-Smith
22 Aug 2011
Simon McIntosh-Smith will be co-running an all-day workshop at the IEEE/ACM Conference on High Performance Computing, Networking, Storage and Analysis (SuperComputing) with Tim Mattson from Intel and Ben Gaster from AMD... read more.

Older news...

Recent publications

Towards Safe Human-Robot Interaction
Elena Corina Grigore, Kerstin Eder, Alexander Lenz, Sergey Skachek, Anthony G. Pipe and Christopher Melhuish, 2011

Upcoming events

The Multicore Challenge II: Programming Multicore Systems
5 Sep 2011 at 1:00 in University of the West of England, Frenchay Campus, Bristol
Experts in multicore technology are coming together in Bristol in September to look at the challenges of developing multicore systems... read more.

More events...

Supporters & affiliations

The μ Research Group works closely with the following companies and organisations.