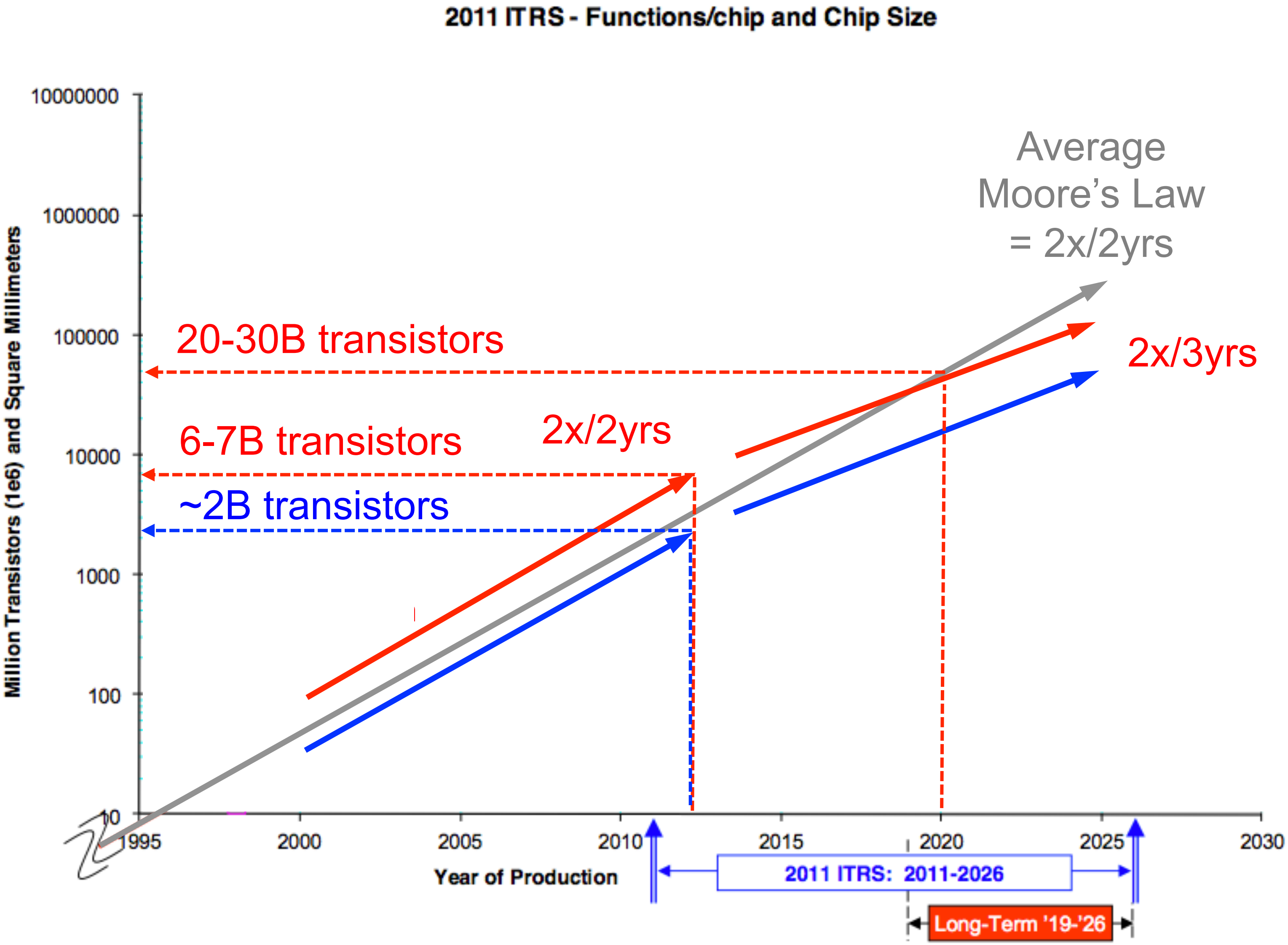
Trends in Heterogeneous Systems Architectures (and how they'll affect parallel programming models)

Simon McIntosh-Smith simonm@cs.bris.ac.uk
Head of Microelectronics Research
University of Bristol, UK

University of BRISTOL

Moore’s Law today



http://www.itrs.net/Links/2011ITRS/2011Chapters/2011ExecSum.pdf

Herb Sutter’s new outlook

http://herbsutter.com/welcome-to-the-jungle/

“In the twilight of Moore’s Law, the transitions to multicore processors, GPU computing, and HaaS cloud computing are not separate trends, but aspects of a single trend – mainstream computers from desktops to ‘smartphones’ are being permanently transformed into heterogeneous supercomputer clusters. Henceforth, a single compute-intensive application will **need to harness different kinds of cores, in immense numbers**, to get its job done.”

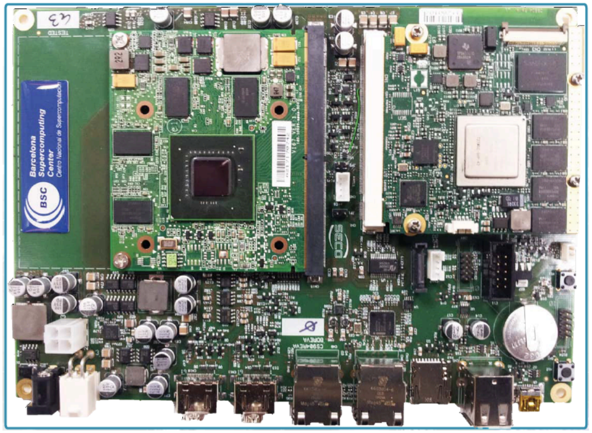
“The free lunch is over.
Now welcome to the ***hardware jungle***.”

Four causes of heterogeneity

- Multiple types of programmable core
  - CPU (lightweight, heavyweight)
  - GPU
  - Others (accelerators, …)
- Interconnect asymmetry
- Memory hierarchies
- Software (OS, middleware, tools, …)

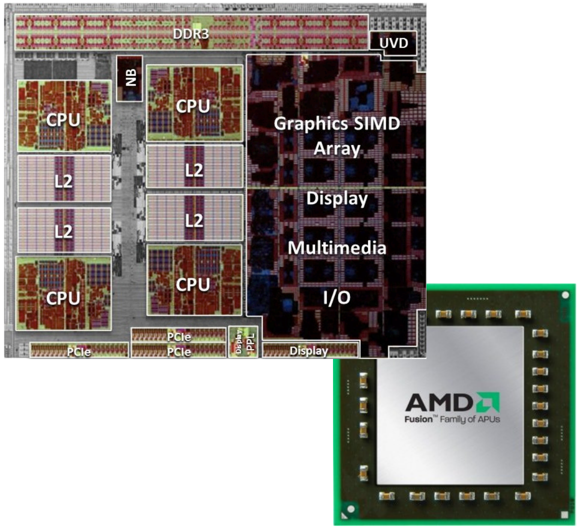
Heterogeneous Systems

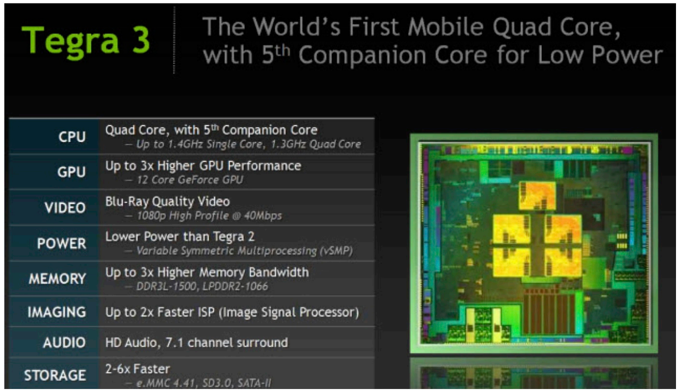
Intel MIC



FP7 Mont Blanc ARM + GPU

AMD Llano Fusion APUs





Tegra 3 The World's First Mobile Quad Core, with 5th Companion Core for Low Power

CPU	Quad Core, with 5th Companion Core – Up to 1.4GHz Single Core, 1.3GHz Quad Core
GPU	Up to 3x Higher GPU Performance – 12 Core GeForce GPU
VIDEO	Blu-Ray Quality Video – 1080p High Profile @ 40Mbps
POWER	Lower Power than Tegra 2 – Variable Symmetric Multiprocessing (vSMP)
MEMORY	Up to 3x Higher Memory Bandwidth – DDR3L-1500, LPDDR2-1066
IMAGING	Up to 2x Faster ISP (Image Signal Processor)
AUDIO	HD Audio, 7.1 channel surround
STORAGE	2-6x Faster – e.MMC 4.41, SD3.0, SATA-II

NVIDIA Tegra, Project Denver

Heterogeneity is mainstream

Quad-core ARM Cortex A9 CPU
Quad-core SGX543MP4+ Imagination GPU

Dual-core ARM 1.4GHz, ARMv7s CPU
Triple-core SGX554MP4 Imagination GPU

Most tablets and smartphones are already powered by heterogeneous processors.

Current limitations

- Disjoint view of memory spaces between CPUs and GPUs
- Hard partition between “host” and “devices” in programming models
- Dynamically varying nested parallelism almost impossible to support
- Large overheads in scheduling heterogeneous, parallel tasks

University of BRISTOL

The emerging Heterogeneous System Architecture (HSA) standard

Current HSA members

Founders

AMD ARM Imagination

MEDIATEK QUALCOMM® SAMSUNG TEXAS INSTRUMENTS

Promoters

LG Electronics

Supporters

Arteris codeplay® FABRICENGINE MULTICORE WARE

Contributors

apical CEVA DMP MARVELL ST life.augmented

Sony Mobile ST ERICSSON SONICS <symbio> tensilica VIVANTE

Academic

NTHU Programming Language Lab

University of Illinois Computer Science

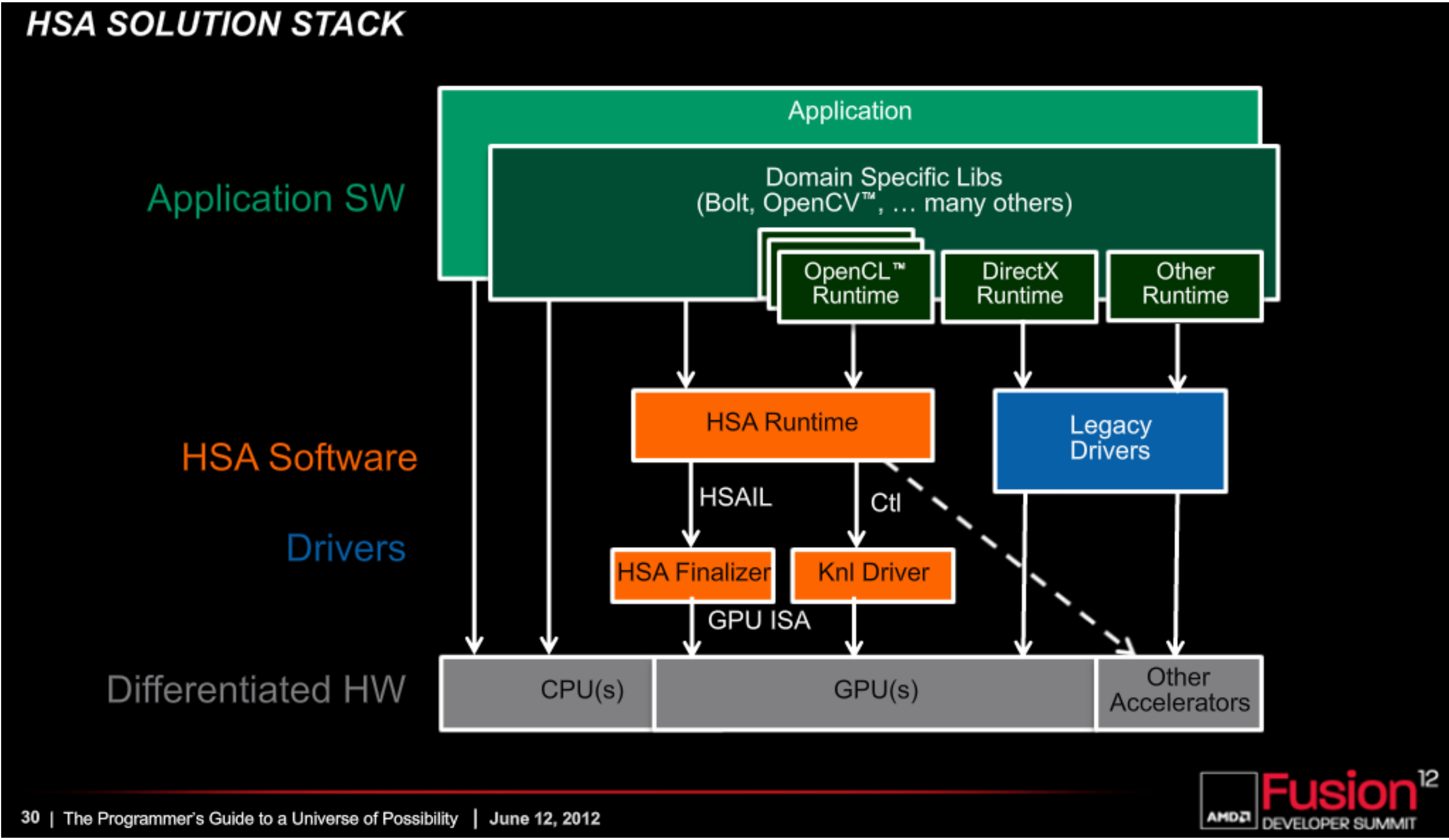
THE UNIVERSITY of EDINBURGH School of informatics

University of BRISTOL

HSA overview

- The HSA Foundation launched mid 2012
- HSA is a new, ***open*** architecture specification
  - HSAIL virtual (parallel) instruction set
  - HSA memory model
  - HSA dispatcher and run-time
- Provides an optimised platform architecture for heterogeneous programming models such as OpenCL, C++AMP, et al

HSA overview



HSA SOLUTION STACK

Application SW

HSA Software

Drivers

Differentiated HW

Application

Domain Specific Libs
(Bolt, OpenCV™, … many others)

OpenCL™ Runtime

DirectX Runtime

Other Runtime

HSA Runtime

Legacy Drivers

HSAIL

Ctl

HSA Finalizer

Knl Driver

GPU ISA

CPU(s)

GPU(s)

Other Accelerators

30 | The Programmer's Guide to a Universe of Possibility | June 12, 2012

AMD Fusion12 DEVELOPER SUMMIT

Enabling more efficient heterogeneous programming

- Unified virtual address space for all cores
  - CPU and GPU
  - Enables PGAS-style distributed arrays
- Hardware queues per core with lightweight user mode task dispatch
  - Enables GPU context switching, preemption, efficient heterogeneous scheduling
- First class barrier objects
  - Aids parallel program composability

HSA Intermediate Layer (HSAIL)

- Virtual ISA for parallel programs
- Similar to LLVM IR and OpenCL SPIR
- *Finalised* to specific ISA by a JIT compiler
- Make late decisions on which core should run a task
- HSAIL features:
  - Explicitly parallel
  - Support for exceptions, virtual functions and other high-level features
  - Syscall methods (I/O, printf etc.)
  - Debugging support

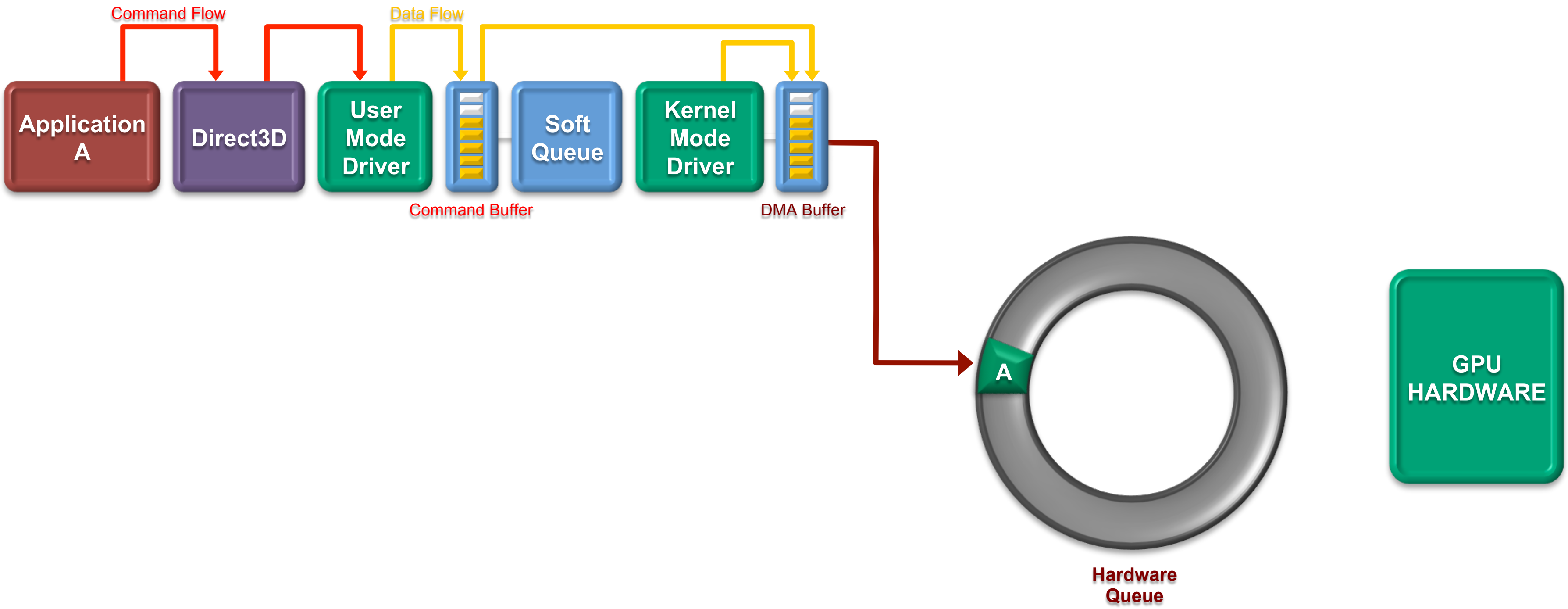
HSA memory model

- Compatible with C++11, OpenCL, Java and .NET memory models
- Relaxed consistency
- Designed to support both managed language (such as Java) and unmanaged languages (such as C)
- Will make it much easier to develop 3rd party compilers for a wide range of heterogeneous products
  - E.g. Fortran, C++, C++AMP, Java et al

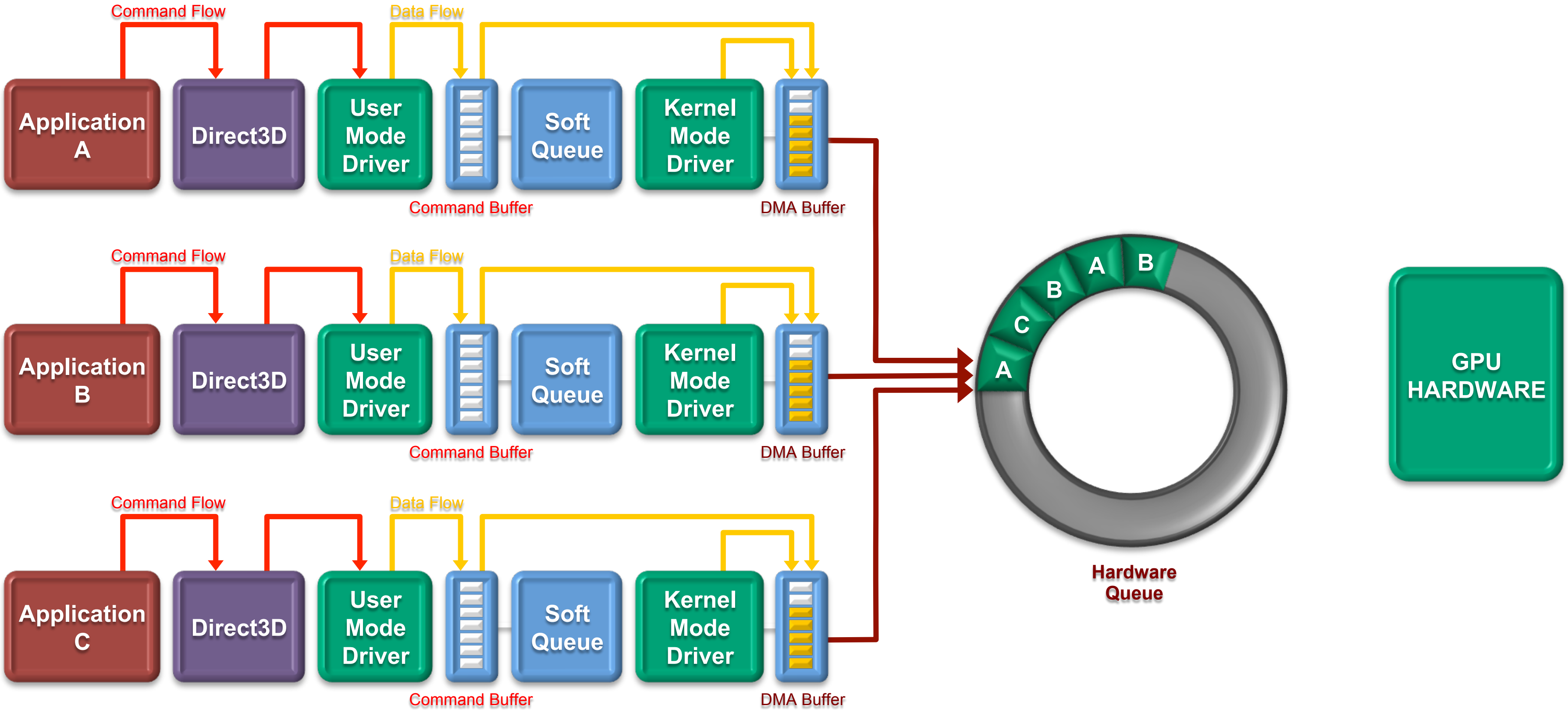
HSA dispatch

- HSA designed to enable heterogeneous task queuing
  - A work queue per core (CPU, GPU, …)
  - Distribution of work into queues
  - Load balancing by work stealing
- Any core can schedule work for any other, including itself
- Significant reduction in overhead of scheduling work for a core

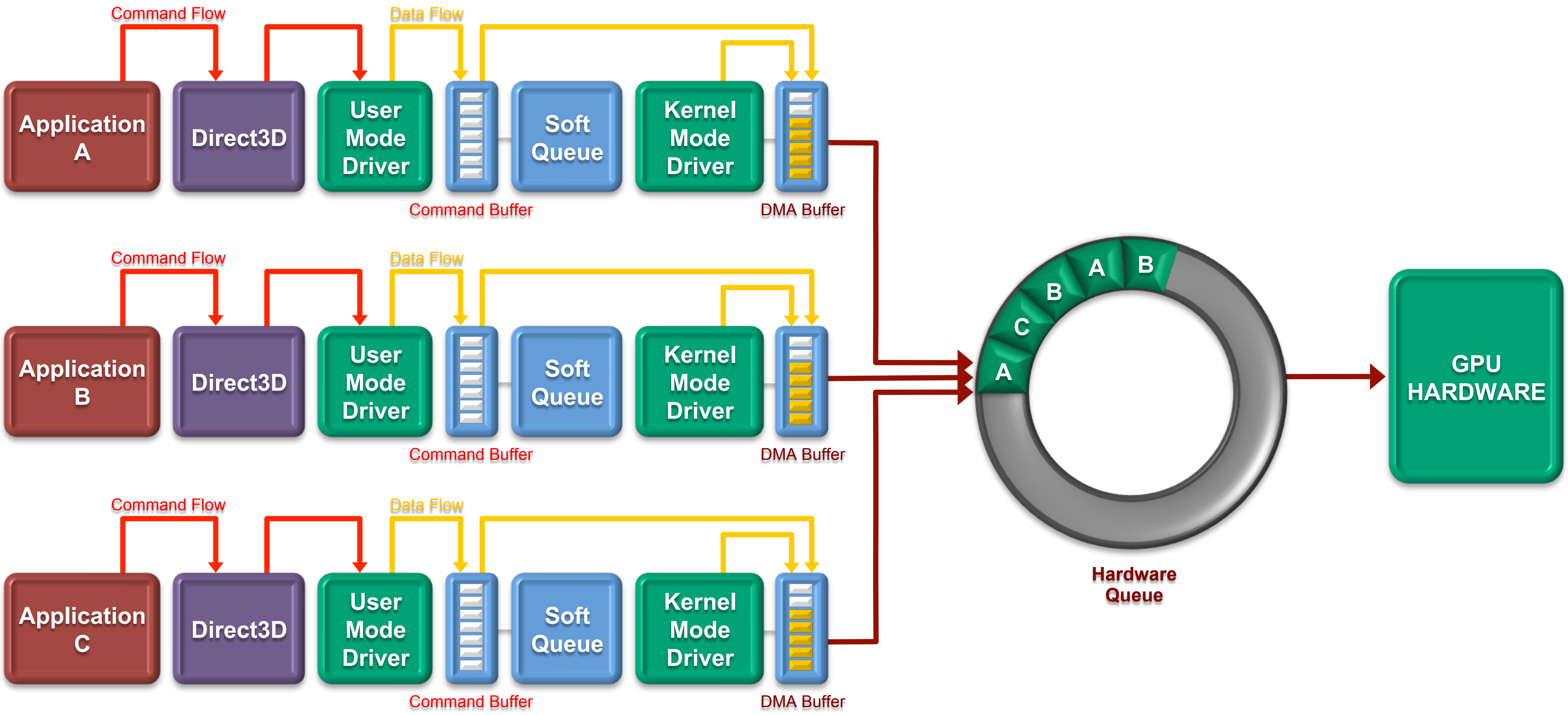
Today's Command and Dispatch Flow



Today’s Command and Dispatch Flow

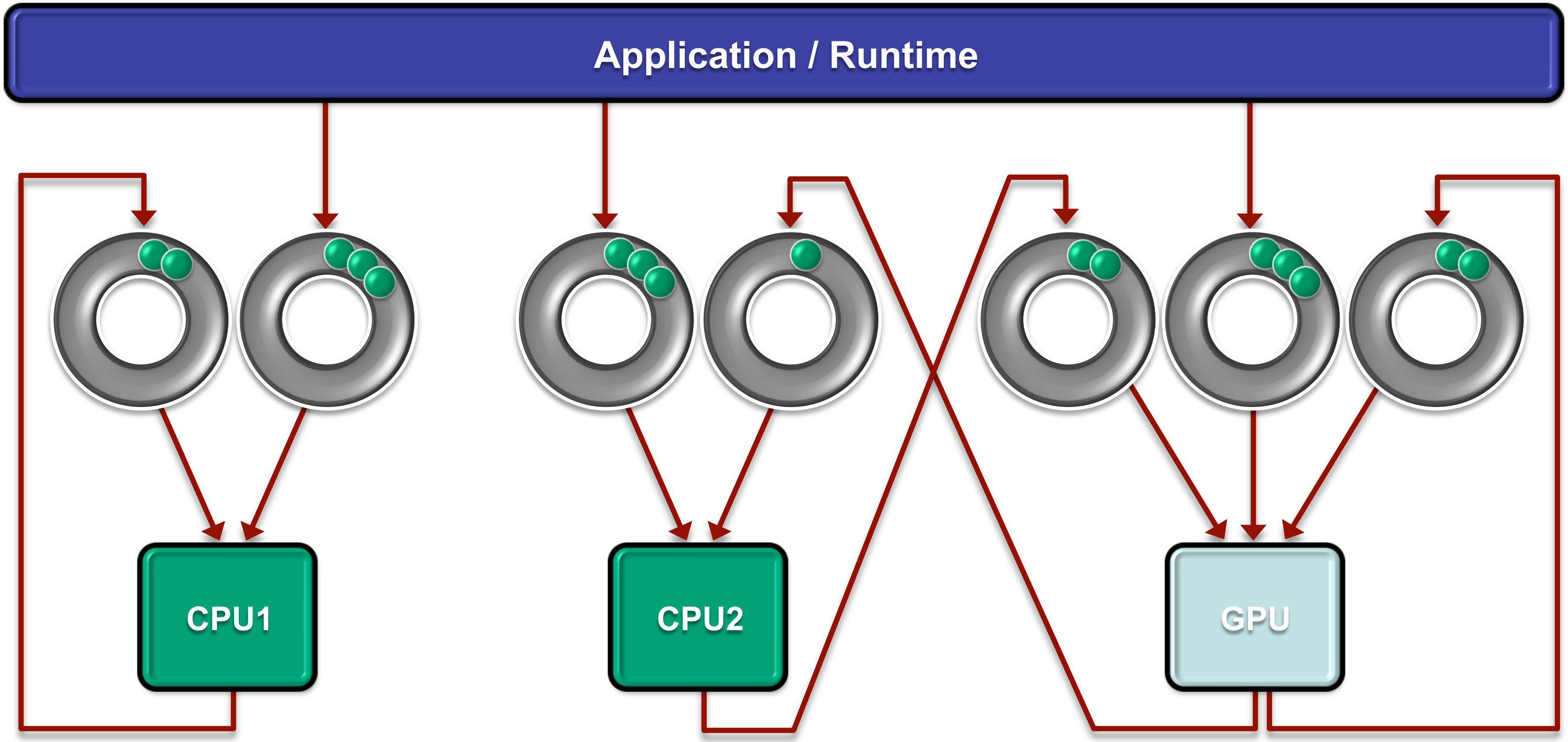


Today's Command and Dispatch Flow

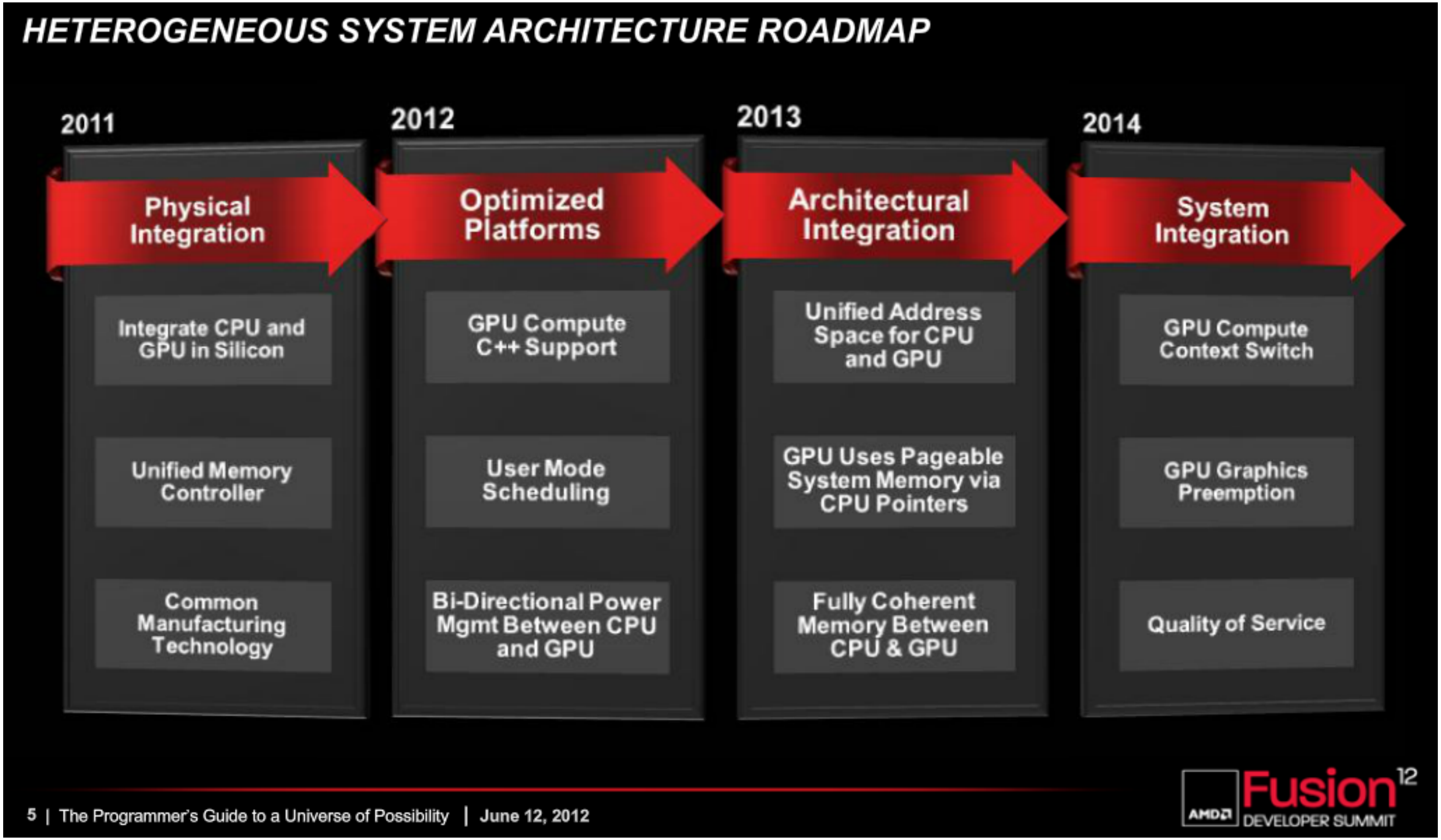


University of BRISTOL

HSA enabled dispatch



HSA roadmap from AMD



HETEROGENEOUS SYSTEM ARCHITECTURE ROADMAP

2011	2012	2013	2014
Physical Integration	Optimized Platforms	Architectural Integration	System Integration
Integrate CPU and GPU in Silicon	GPU Compute C++ Support	Unified Address Space for CPU and GPU	GPU Compute Context Switch
Unified Memory Controller	User Mode Scheduling	GPU Uses Pageable System Memory via CPU Pointers	GPU Graphics Preemption
Common Manufacturing Technology	Bi-Directional Power Mgmt Between CPU and GPU	Fully Coherent Memory Between CPU & GPU	Quality of Service

5 | The Programmer's Guide to a Universe of Possibility | June 12, 2012

AMD Fusion12 DEVELOPER SUMMIT

Open Source software stack for HSA

A Linux execution and compilation stack will be open-sourced by AMD

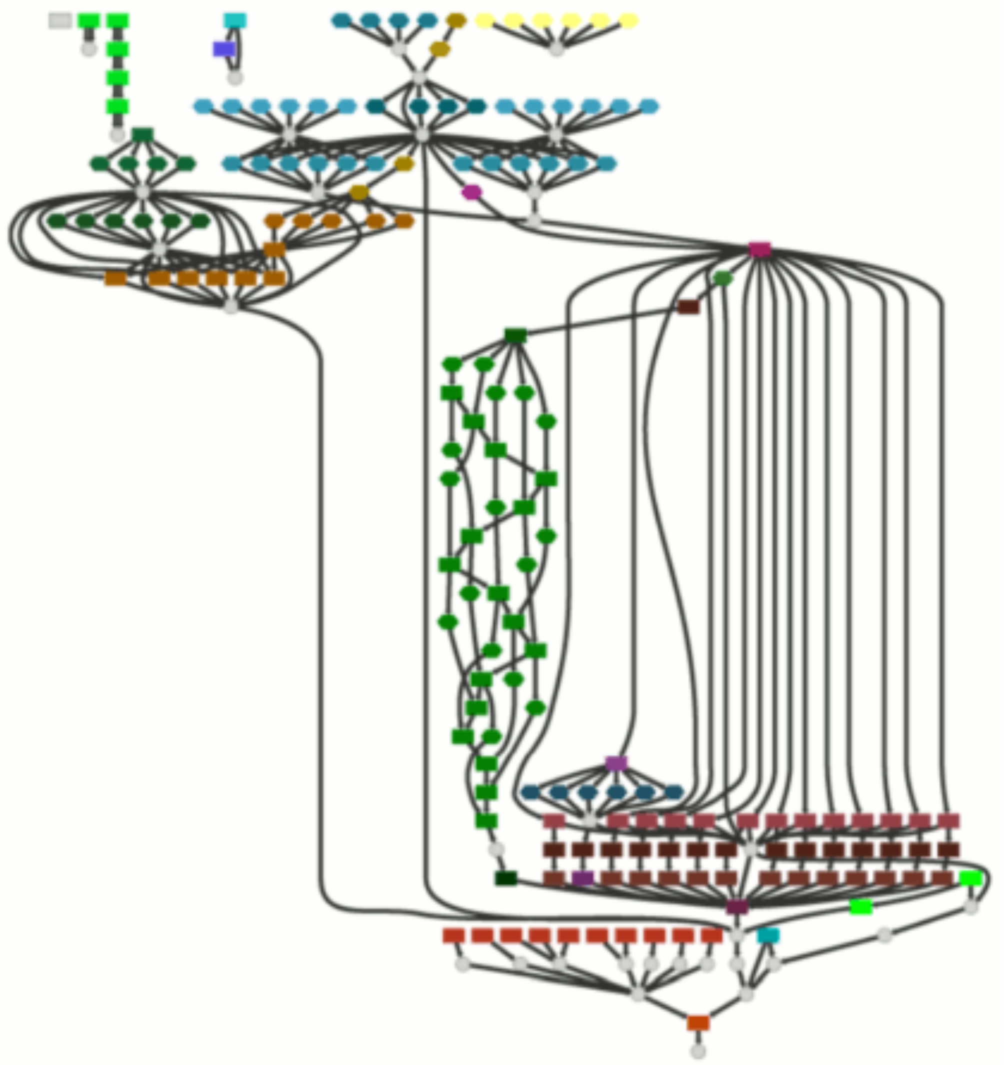
- Jump start the ecosystem
- Allow a single shared implementation where appropriate
- Enable university research in all areas

Component Name	Purpose
HSA Bolt Library	Enable understanding and debug
OpenCL HSAIL Code Generator	Enable research
LLVM Contributions	Industry and academic collaboration
HSA Assembler	Enable understanding and debug
HSA Runtime	Standardize on a single runtime
HSA Finalizer	Enable research and debug
HSA Kernel Driver	For inclusion in Linux distros

HSA should enable nested parallel programs like this

Support for multiple algorithms, even within a single application

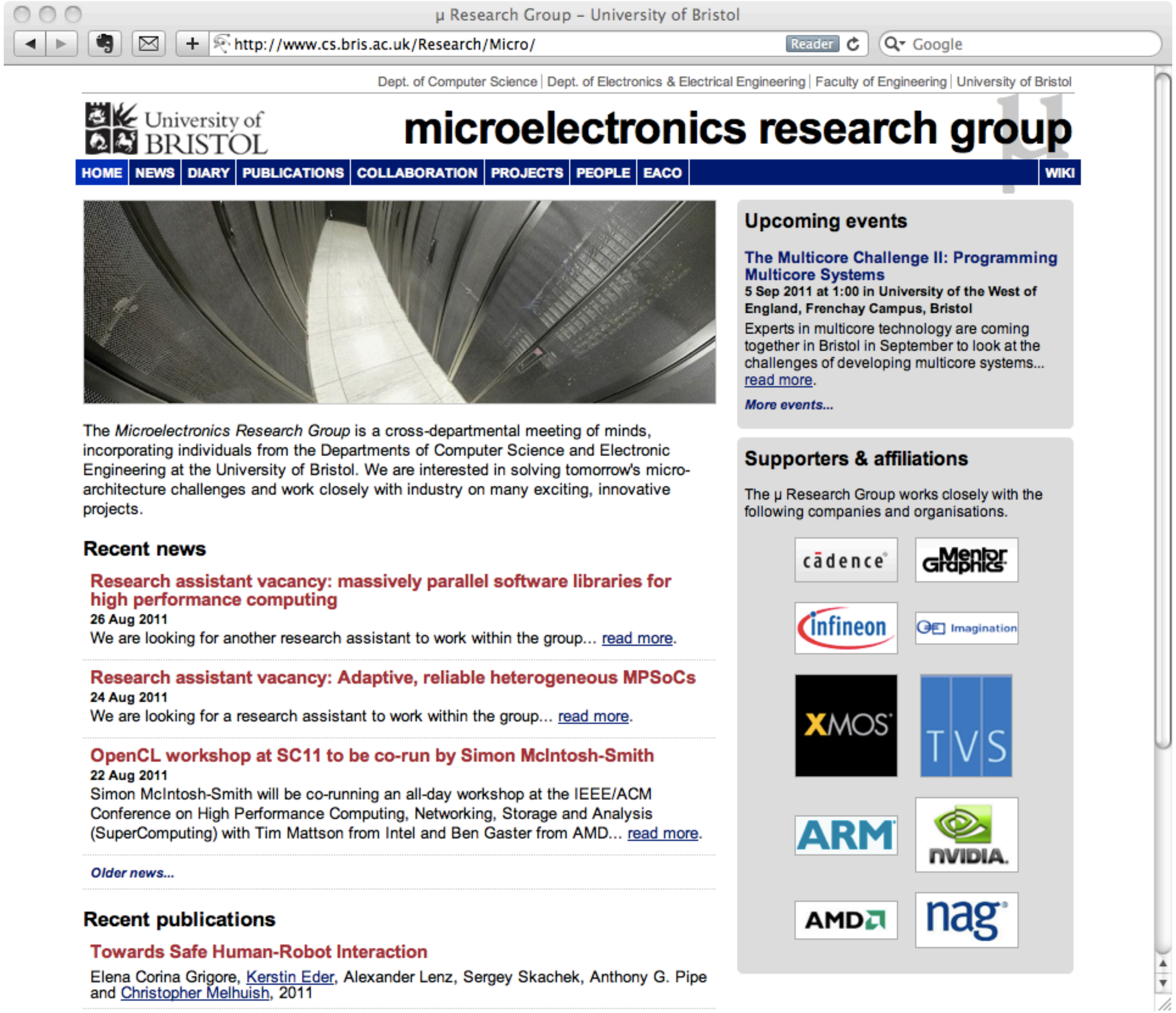
Task farms, pipeline, data parallelism, …



Conclusions

- Heterogeneity is an increasingly important trend
- The market is finally starting to create and adopt the necessary open standards
  - Proprietary models likely to start declining now
  - Don’t get locked into any one vendor!
- Parallel programming models are likely to (re)proliferate
- HSA should enable much more dynamically heterogeneous nested parallel programs and programming models

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.