Rechnerarchitektur (RA)

Sommersemester 2017

Architecture-Aware Optimizations
- Hardware-Software co-Optimizations-

Jian-Jia Chen
Informatik 12
Jian-jia.chen@tu-..
Tel.: 0231 755 6078

2017/07/25
Outline

Heterogeneous System Architecture (HSA)
- Integrated CPU/GPU platforms
- Recent movement in chip designs

Architecture-aware software designs
- Energy-efficiency issues
- Darkroom
- Halide

Multicore revolutions
- Impact on the “safety-critical” industry sector
What is Heterogeneous Computing?

Use processor cores with various type/computing power to achieve better performance/power efficiency

Advantage of Heterogeneous Computing

CPU is ideal for scalar processing
- Out of order x86 cores with low latency memory access
- Optimized for sequential and branching algorithms
- Runs existing applications very well

Serial/Task-parallel workloads → CPU

GPU is ideal for parallel processing
- GPU shaders optimized for throughput computing
- Ready for emerging workloads
- Media processing, simulation, natural UI, etc.

Graphics/Data-parallel workloads → GPU

Heterogeneous Computing -> Fusion, Norm Rubin, SAAHPC 2010
CPU/GPU Integration: CPU’s Advancement Meets GPU’s

Microprocessor Advancement

- Single-Thread Era
- Multi-Core Era
- Heterogeneous Systems Era

High Performance Task Parallel Execution

System-Level Programmable

Vertex/Pixel Shader Graphics Driver-based programs

Power-efficient Data Parallel Execution

Unacceptable

Experts Only

Mainstream

Programmability

Throughput Performance

GPU Advancement
Evolution of Heterogeneous Computing

**Dedicated GPU**
- GPU kernel is launched through the device driver
- Separate CPU/GPU address space
- Separate system/GPU memory
- Data copy between CPU/GPU via PCIe

---

**OpenCL Application**

**OpenCL Runtime Library**

**User Space**

**Kernel Space**

**GPU Device Driver**

**Address space managed by OS**

**Address space managed by driver**

---

**System memory**

(coherent)

---

**GPU memory**

(Non-coherent)
Evolution of Heterogeneous Computing

Integrated GPU architecture
- GPU kernel is launched through the device driver
- Separate CPU/GPU address space
- Separate system/GPU memory
- Data copy between CPU/GPU *via memory bus*

![Diagram of GPU architecture](image)
Evolution of Heterogeneous Computing

Integrated CPU/GPU architecture

- GPU kernel is launched through the device driver
- Unified CPU/GPU address space (managed by OS)
- Unified system/GPU memory
- No data copy - data can be retrieved by pointer passing

![Diagram showing the evolution of heterogeneous computing with integrated CPU/GPU architecture.](image-url)
Utopia World of Heterogeneous Computing

Processors are architected to operate cooperatively
- Tasks in an application are executed on different types of core
- Unified coherent memory enables data sharing across all processors

Designed to enable the applications to run on different processors at different time
- Capability to translate from high-level language to target binary at run-time
- User-level task dispatch
- Decision making module

![Diagram of heterogeneous computing system](image)
HSA Foundation

Founded in June 2012
Developing a new platform for heterogeneous systems
www.hsafoundation.com
Specifications under development in working groups to define the platform
Membership consists of 43 companies and 16 universities
Adding 1-2 new members each month
Diverse Partners Driving Future of Heterogeneous Computing

Founders

Promoters

Supporters

Contributors
HSA (Heterogeneous System Architecture)

HSA Run-time

Application

OpenCL

User-level SW

HW

Native machine codes

CPU Finalizer

GPU Finalizer

Agent Scheduler

Core 1

Core 2

CU 1

CU 2

AC 1

AC 2

Shared virtual address space
Intel Haswell

Haswell Processor Family Overview (Traditional)

- 22nm Process
- Socket: G3 (947 pin) MB
  H3 (1150 pin) DT
- Fully Integrated VR
- Power Aware Interrupt Routing for power / performance
- Intel Hyper-Threading Technology
- Intel AVX 2.0 extensions and AES-NI Instructions Improvements
- Intel Turbo Boost Technology
- Last Level Cache (LLC) shared between CPU and Integrated Graphics
- Hotham 1.0 Support
- Overclocking Improvements

- Processor Graphics
  - 3D Graphics/Media Arch/Perf.
  - Support for new APIs
  - 3 Digital Displays
  - DP 1.2/HDMI 1.4/eDP
- Desktop/All-In-One
  - DDR3/DDR3L-1600 2D/Ch
  - non-ECC UDIMM/SODIMM
- Mobile
  - DDR3L (only)
  - POR for 1600 2D/Ch 6L rPGA
  - non-ECC SODIMM
- DDR Power Gating for lower idle power (MB)
- PCIe 3.0 x16/2x8
  - 3 controllers
- Power Optimizer (CPMP) Support (MB)

Haswell Offers Better Power & Performance with Improved I/O
**AMD Kaveri APU**

“KAVERI” FEATURING UP TO 12 COMPUTE CORES (4 CPU+ 8 GPU)

---

**CPU COMPUTE CORES**

Up to four new multi-threaded AMD “Steamroller” CPU CORES

---

**GPU COMPUTE CORES**

Up to eight GCN GPU CORES powering parallel compute and next-gen gaming

---

Visit amd.com/computecores for more details
NVIDIA Tegra K1

Tegra K1

Kepler GPU (192 CUDA Cores)
Open GL 4.4, OpenGL ES3.0, DX11, CUDA 6

Quad Core Cortex A15 “r3”
With 5th Battery-Saver Core; 2MB L2 cache

Dual High Performance ISP
1.2 Gigapixel throughput, 100MP sensor

Lower Power
28HPM, Battery Saver Core

4K panel, 4K HDMI
DSI, eDP, LVDS, High Speed HDMI 1.4a
Qualcomm Snapdragon

- Location
  - GPS, GLONASS, Beidou, Galileo Satellites

- Adreno 430 GPU
  - OpenGL ES 2.0/3.1
  - OpenCL 1.2 Full
  - Content Security

- Cortex-A57 & Cortex-A53 CPUs

- Memory
  - LPDDR4

- Display Processing
  - 4K, Miracast, picture enhancement

- Hexagon DSP
  - Ultra Low Power
  - Sensor Engine

- Modem
  - 4th gen CAT 6 LTE
  - Up to 3x20MHz CA

- USB 3.0

- Multimedia Processing
  - 4K Encode/Decode
  - Snapdragon Voice Activation
  - Gestures
  - Studio Access Security

- Dual ISPs (Camera)
  - Up to 55MP
  - 1.2Gpix/s bw
  - Camera SW
Outline

Heterogeneous System Architecture (HSA)
- Integrated CPU/GPU platforms
- Recent movement in chip designs

Architecture-aware software designs
- Energy-efficiency issues
- Darkroom
- Halide

Multicore revolutions
- Impact on the “safety-critical” industry sector
Energy Efficiency of different target platforms

© Hugo De Man, IMEC, Philips, 2007
Signal Processing ASICs

Markovic, EE292 Class, Stanford, 2013
How about Memory?
Processor Energy Breakdown

- 8 cores
- L1/reg/TLB
- L2
- L3

© Horowitz, DAC 2016
Data Center Energy Specs

© Malladi, ISCA 2012
ITRS Power Consumption Projection
-Station Systems -
What Is Going On Here?

Markovic, EE292 Class, Stanford, 2013
## Energy Consumption (Approximate, 45nm)

<table>
<thead>
<tr>
<th>Integer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td></td>
</tr>
<tr>
<td>8 bit</td>
<td>0.03pJ</td>
</tr>
<tr>
<td>32 bit</td>
<td>0.1pJ</td>
</tr>
<tr>
<td>Mult</td>
<td></td>
</tr>
<tr>
<td>8 bit</td>
<td>0.2pJ</td>
</tr>
<tr>
<td>32 bit</td>
<td>3 pJ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FAdd</td>
<td></td>
</tr>
<tr>
<td>16 bit</td>
<td>0.4pJ</td>
</tr>
<tr>
<td>32 bit</td>
<td>0.9pJ</td>
</tr>
<tr>
<td>FMult</td>
<td></td>
</tr>
<tr>
<td>16 bit</td>
<td>1pJ</td>
</tr>
<tr>
<td>32 bit</td>
<td>4pJ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cache</td>
<td>(64bit)</td>
</tr>
<tr>
<td>8KB</td>
<td>10pJ</td>
</tr>
<tr>
<td>32KB</td>
<td>20pJ</td>
</tr>
<tr>
<td>1MB</td>
<td>100pJ</td>
</tr>
<tr>
<td>DRAM</td>
<td>1.3-2.6nJ</td>
</tr>
</tbody>
</table>

### Instruction Energy Breakdown

- **I-Cache Access**: 25pJ
- **Register File Access**: 6pJ
- **Add**: 70pJ

© Horowitz, DAC 2016
True Story

It’s more about the algorithm than the hardware
The efficiency cannot be achieved unless the algorithm is right!!

(all) Algorithms

GPU Alg.
Locality, Locality, and Locality!!!

© Hegarty et al., SIGGraph 2014
Darkroom (Stanford/MIT)

\[
\begin{align*}
bx & = im(x,y) \\
& \quad (I(x-1,y) + I(x,y) + I(x+1,y))/3 \\
n & = \text{by} = im(x,y) \\
& \quad (bx(x,y-1) + bx(x,y) + bx(x,y+1))/3 \\
end & \\
\text{sharpened} & = im(x,y) \\
& \quad I(x,y) + 0.1 \ast (I(x,y) - \text{by}(x,y)) \\
end & \\
\end{align*}
\]

© Hegarty et al., SIGGraph 2014
Locality versus Parallelism

Halide Programming Language:
- http://halide-lang.org/

Performance needs a lot of tradeoffs
- Locality
- Parallelism
- Redundant recomputation
Outline

Heterogeneous System Architecture (HSA)
- Integrated CPU/GPU platforms
- Recent movement in chip designs

Architecture-aware software designs
- Energy-efficiency issues
- Darkroom
- Halide

Multicore revolutions
- Impact on the “safety-critical” industry sector
Automotive Software

**Task**

- **Activation Pattern:**
  - Periodic: 1 to 1000 ms
  - Angle synchronous
  - Sporadic

- **Scheduled by the OS**
  - Fixed Priorities
  - Preemptively or cooperatively

**OSEK task states**

- **Extended Task**
  - States: **running**, **waiting**
  - Transitions: start, preempt, terminate

- **Basic Task**
  - States: **suspended**, **ready**
  - Transitions: activate, release

© Hamann, Kramer, Ziegenbein, Lukasiewycz (Bosch), 2016
Assessment of Multi-Core Worst-Case Execution Behavior

Communication between runnables is realized with reading and writing of labels

<table>
<thead>
<tr>
<th>Communication</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>25 %</td>
</tr>
<tr>
<td>Backward</td>
<td>35 %</td>
</tr>
<tr>
<td>InterTask</td>
<td>40 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic (1-4 bytes)</td>
<td>97 %</td>
</tr>
<tr>
<td>Structs / Arrays</td>
<td>3 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Access type</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read-only</td>
<td>40 %</td>
</tr>
<tr>
<td>Write-only</td>
<td>10 %</td>
</tr>
<tr>
<td>Read-Write</td>
<td>50 %</td>
</tr>
</tbody>
</table>

© Hamann, Kramer, Ziegenbein, Lukasiewycz (Bosch), 2016
Multi-Core Memory Access Models

Access time to data in different memories (local & global)

© Hamann, Kramer, Ziegenbein, Lukasiewycz (Bosch), 2016
An Industrial Challenge (FMTV 2016)

Precise analysis of worst-case end-to-end latencies

- mainly due to different involved periods and time domains
- What is the effect on memory layout and interconnect on the execution times?
- Automatic optimized application and data mapping
- Evaluation of digital (multi-core) execution platforms
- Evaluation of software growth scenarios

© Hamann, Kramer, Ziegenbein, Lukasiewycz (Bosch), 2016
MPPA-256 Processor Architecture (Kalray)

Kalray, 2016
MPPA-256 NoC

- Dual 2D-torus NoC
  - D-NoC: high-bandwidth RDMA
  - C-NoC: low-latency mailboxes
  - 4B/cycle per link direction per NoC
  - Nx10Gb/s NoC extensions for connection to FPGA or other MPPA®

- Predictability
  - Data NoC is configured by selecting routes and injection parameters
  - Injection parameters are the $(\sigma, \rho)$ or (burst, rate) of Cruz network calculus
  - Guaranteed services rely on same methods as in AFDX Ethernet
Safety-Critical Systems with Multicore Platforms

**Goal:** deploy multi-core processors for safety-critical real-time applications (avionics, automotive, ...)

**Problem:** concurrent use of shared resources (e.g. interconnect, main memory)
- unknown access latency for a concrete resource access
- complicated timing analysis
- hardware platforms may not be predictable
- Many features are designed by computer architects for *average cases* only

**Solution?**
- Maybe it is up to you.
- Did you see the above challenges?