Rechnerarchitektur
Sommersemester 2013

Manycore-Beschleuniger:
Intel Xeon Phi

Michael Engel
Informatik 12
TU Dortmund
Trends...

- Beschleuniger in normalen Systemen (z.B. GPUs)
  - Hohe Rechenleistung pro Volumen
  - Hohe MFLOPS/W
Historie der Intel-Manycores

- Polaris: 80 Core Experimental Multicore (2007)
- Larrabee
  - Als GPU entworfen
  - 2010
- SCC (2011)
  - Single Chip Cloud Comp.
- MIC-Architektur
  - Knights Bridge (2013)
  - Knights Landing
Historie der Intel-Manycores

- **Polaris Experimental Multicore: Teraflop Research Chip**
  - 80 einfachste Cores @3,16 GHz: 1 TFLOP, 62W
- **Larrabee: Hybrid between a multi-core CPU and a GPU**
  - Coherent cache hierarchy and x86 architecture
  - Wide SIMD vector units, texture sampling hardware
- **SCC**
  - 48 P54C Pentium cores connected by 4x6 2D-mesh
  - Four DDR3 memory controllers on the chip
    - Connected to the 2D-mesh as well
  - Cores are divided into 24 tiles
    - Each tile has 2 cores and a message passing buffer (MPB) shared by the two cores
Intel Polaris: Teraflops Research Chip

- Technology: 65nm CMOS Process
- Interconnect: 1 poly, 8 metal (Cu)
- Transistors: 100 Million
- Die Area: 275mm²
- Tile area: 3mm²
- Package: 1248 pin LGA, 14 layers, 343 signal pins
Intel Polaris: Teraflops Research Chip (2)

...erinnert an Connection Machine
**Tiled Design & Mesh Network**

**Repeated Tile Method:**
- Compute + router
- Modular, scalable
- Small design teams
- Short design cycle

**Mesh Interconnect:**
- “Network-on-a-Chip”
  - Cores networked in a grid allows for super high bandwidth communications in and between cores
- 5-port, 80GB/s* routers
- Low latency (1.25ns*)
- Future: connect IA/or and special purpose cores

* When operating at a nominal speed of 4GHz

Content under media embargo through Sunday, February 11th Noon PST
Intel Polaris: Energie

Fine Grain Power Management

- Novel, modular clocking scheme saves power over global clock
- New instructions to make any core sleep or wake as apps demand
- Chip Voltage & freq. control (0.7-1.3V, 0-5.8GHz)

Dynamic sleep

STANDBY:
- Memory retains data
- 50% less power/tile

FULL SLEEP:
- Memories fully off
- 80% less power/tile

21 sleep regions per tile (not all shown)

Data Memory
- Sleeping: 57% less power

Instruction Memory
- Sleeping: 56% less power

Router
- Sleeping: 10% less power (stays on to pass traffic)

FP Engine 1
- Sleeping: 90% less power

FP Engine 2
- Sleeping: 90% less power

Industry leading energy-efficiency of 16 Gigaflops/Watt
**Research Data Summary**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Voltage</th>
<th>Power</th>
<th>Bisection Bandwidth</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.16 GHz</td>
<td>0.95 V</td>
<td>62W</td>
<td>1.62 Terabits/s</td>
<td>1.01 Teraflops</td>
</tr>
<tr>
<td>5.1 GHz</td>
<td>1.2 V</td>
<td>175W</td>
<td>2.61 Terabits/s</td>
<td>1.63 Teraflops</td>
</tr>
<tr>
<td>5.7 GHz</td>
<td>1.35 V</td>
<td>265W</td>
<td>2.92 Terabits/s</td>
<td>1.81 Teraflops</td>
</tr>
</tbody>
</table>

1.01 Teraflops  
62 Watts
Xeon Phi

- Beschleuniger für HPC
Flashback: TOP500 Supercomputer-Liste

- Mai 2013 – Platz #1: Tianhe-2
  - 33,86 Petaflops
- 16.000 compute nodes bestehend aus:
  - Zwei Intel Ivy Bridge Xeon 12-Core CPUs
  - Drei Xeon Phi Chips (je 60 Cores)
  - Insgesamt 3.120.000 Cores
- Leistungsaufnahme 17,6 MW
  - 24 MW mit Klimatisierung
- 1.375 TiB RAM
- Interconnect: fat tree-Topologie
  - 13 switches each of 576 ports
- 12,4 PB Massenspeicher
- Kosten: ca. 390 Millionen US$
Xeon Phi Die Shot

- 60 Cores – ein Die
Xeon Phi

- Identischer Code auf Host und Beschleuniger?
- Fast – Cross-Compiler erforderlich
- Aber: gleiche Compiler (gcc, icc) verwendbar
Xeon Phi-Platine

- PCIe x 16-Karte: 60-Core Xeon Phi, 8 GB GDDR5 RAM
Xeon Phi-Platine: Struktur
Xeon Phi

- x86-basierter SMP-on-a-chip
  - In-order, dual issue Prozessoren
  - Verwandt mit ursprünglichem Pentium (1996!)
- >= 50 cores @ >=1GHz, in 22nm-Technologie gefertigt

- Unterschiede zum Pentium:
  - 64 bit-Support
  - Hardware Multithreading: 4 Threads/Core
  - 512 bit SIMD-Befehle
  - Power Management
  - Ring Interconnect

- CPU-Cores: <2% der Xeon Phi-Chipfläche (ohne L2$)
Xeon Phi: Architektur
Xeon Phi: Interconnect

- Cores interconnected by high-speed bidirectional ring
Xeon Phi: Caches

- Cache coherent **across the entire coprocessor**
- Global distributed tag directory
  - When core accesses its L2 cache and misses, address request sent on address ring to tag dirs
    - Memory addresses uniformly distributed amongst tag dirs on the ring
  - Requested data block found in another core’s L2$:$
    - Forwarding request sent to that L2 over the address ring
    - Request block subsequently forwarded on data block ring
  - Otherwise, memory address is sent from the tag directory to the memory controller
- Each core has a 512-KB L2 cache locally with high-speed access to all other L2 caches
  - making the collective L2 cache size over 25 MB
Caches: Distributed Tag Directory

Tag Directories track cache-lines in all L2s
Coherent Caches are a key MIC Architecture Advantage

- high data BW
- low energy per byte of data supplied
- programmer friendly (coherence just works)
Cache-Nutzung: Stencils

- Spatial time-step simulation of a physical system
- Cache blocking promotes much higher performance and performance/Watt vs. memory streaming
Xeon Phi: Speicherzugriff – Streaming Stores

- Significantly reduces bw needed for full cache-line stores
  - Vector-aligned unmasked stores in streaming kernels
- Avoid wasting memory bandwidth due to reading original content of entire cache line from memory when overwriting its whole content completely
  - New VMOVNRNGOAPS/VMOVNRNGOAPD insns

**Streams Triad**

\[
\text{for } (i=0; i<\text{HUGE}; i++) \\
A[i] = k*B[i] + C[i];
\]

**Without Streaming Stores**

Read A, B, C, Write A

256 Bytes transferred to/from memory per iteration

**With Streaming Stores**

Read B, C, Write A

192 Bytes transferred to/from memory per iteration
Xeon Phi: Vektoreinheit

Vector ALUs
16 Wide x 32 bit
8 Wide x 64 bit
Fused Multiply Add
Xeon Phi: Vektoreinheit (2)

- 16 wide SP SIMD, 8 wide DP SIMD
- 2:1 Ratio good for circuit optimization
Gather/Scatter Address Machinery

- Gather/Scatter machine takes advantage of cache-line locality
Xeon Phi: Performance – Theoretisches Maximum –

- Xeon E5-2670 vs. Xeon Phi 5110P und SE10P/X
Xeon Phi: Performance – Synthetische Benchmarks –

- Xeon E5-2670 vs. Xeon Phi 5110P und SE10P/X

Coprocessor results: Benchmark run 100% on coprocessor, no help from Intel® Xeon® processor host (aka native)
Xeon Phi:
Performance vs. #Threads

- 1000x1000 Matrix-Multiplikation mit OpenMP
- Peak-Performance
  - Peaks bei 120 und ~240 Threads (2/4x #Cores)
Xeon Phi: Performance vs. #Threads

- 1000x1000 Matrix-Multiplikation mit OpenMP
- Durchschnittliche Performance
  - Teilweise starke Variation der Laufzeiten
  - OS-Einfluß, User-Prozesse, RR-Scheduling
Performance und Leistungsaufnahme

- Performance / Watt (Performance/Coprocessor power)

Vergleich mit Polaris:
16 GFLOP/W!

Measured Coprocessor Power
Power Management

- Intel C-Step modi: aus “normalen” x86-CPUs bekannt
- Clock vs. power gating
  - Package Auto-C3: Ring and Uncore clock gated
  - Package Deep-C3: VccP reduced
  - Package C6: VccP is off (i.e. Cores, Ring and Uncore are powered down)

<table>
<thead>
<tr>
<th>Package Idle State</th>
<th>Core State</th>
<th>Uncore State</th>
<th>TSC/LAPIC</th>
<th>C3WakeupTimer</th>
<th>PCI Express* Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC3</td>
<td>Preserved</td>
<td>Preserved</td>
<td>Frozen</td>
<td>On expiration, package exits PC3</td>
<td>Package exits PC3</td>
</tr>
<tr>
<td>Deep C3</td>
<td>Preserved</td>
<td>Preserved</td>
<td>Frozen</td>
<td>No effect</td>
<td>Times out</td>
</tr>
<tr>
<td>PC6</td>
<td>Lost</td>
<td>Lost</td>
<td>Reset</td>
<td>No effect</td>
<td>Times out</td>
</tr>
</tbody>
</table>
Kooperation von Host und Xeon Phi

- MPSS coprocessor driver
- Coprocessor idle states
  - Core-C0
  - Core-C1
  - Core-C6
  - Package Auto-C3
  - Package Deep-C3
  - Package C6

Shared Coprocessor Power Management
Power Management: Vollbetrieb

- Alle Cores aktiv
Power Management: Clock Gating

- Core C1: Clock Gate Core
- When all 4T on a core have halted, core clock gates itself
Power Management: Power Gating

- Core C6: Power Gate Core
- C1 time-out, power gate core, save leakage, requires core-re-init
Power Management: Package Auto C3

- Timeout when all cores have been in C6
- Clock gate the L2 and interconnect
Power Management: Package C6

- Host Driver can initiate Package C6
- Uncore Voltage Off, requires partial restart
Xeon Phi: System Power Profile (1)

- Alle Angaben in W
- Compute intensive workloads
Xeon Phi: System Power Profile (2)

- Alle Angaben in W
- Memory intensive workloads
Xeon Phi: Kooperation mit Host

- Offload Execution Mode
  - Also known as heterogeneous programming mode
  - Host system offloads part or all of the computation from one or multiple processes running on host

- Coprocessor Native Execution Mode
  - Xeon Phi hosts a Linux micro OS
  - Can appear as another machine connected to the host like another node in a cluster
  - This execution environment allows the users to view the coprocessor as another compute node

- Symmetric Execution
  - Application processes run on both host and Phi
  - They usually communicate through some sort of message passing interface like MPI
Xeon Phi: Kooperation mit Host
Intel Many Integrated Core Platform Software Stack (MPSS) Operating System

61 Cores
Je 4 Threads
Die Zukunft

- Ist leichtere Programmierbarkeit überzeugend?
  - GPUs schwer zu programmieren, aber günstig durch Massenproduktion (→ Gamer 😊)
  - FPGAs schwer zu programmieren und teuer, aber hohe Geschwindigkeit durch angepasste Hardware
Danke!

“The best way to predict the future is to invent it”
– Alan Kay