Heterogeneity Wrap Up
A Case for
Asymmetry Everywhere

Onur Mutlu,
"Asymmetry Everywhere (with Automatic Resource Management)"
Asymmetry Enables Customization

- **Symmetric**: One size fits all
  - Energy and performance suboptimal for different phase behaviors

- **Asymmetric**: Enables tradeoffs and customization
  - Processing requirements vary across applications and phases
  - Execute code on best-fit resources (minimal energy, adequate perf.)
**Thought Experiment: Asymmetry Everywhere**

- Design each hardware resource with **asymmetric, (re-)configurable, partitionable components**
  - Different power/performance/reliability characteristics
  - To fit different computation/access/communication patterns

<table>
<thead>
<tr>
<th>High–power High perf.</th>
<th>Asymmetric / configururable cores and accelerators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power/performance optimized for each access pattern</td>
<td>Asymmetric / partitionable memory hierarchies</td>
</tr>
<tr>
<td>Different technologies Power characteristics</td>
<td>Asymmetric / partitionable interconnect</td>
</tr>
<tr>
<td></td>
<td>Asymmetric main memories</td>
</tr>
</tbody>
</table>
Thought Experiment: Asymmetry Everywhere

- Design the runtime system (HW & SW) to automatically choose the best-fit components for each phase
  - Satisfy performance/SLA with minimal energy
  - Dynamically stitch together the “best-fit” chip for each phase

**Phase 1**
- High-power
- High perf.

**Phase 2**
- Power/performance optimized for each access pattern

**Phase 3**
- Different technologies
- Power characteristics

- Asymmetric / configurable cores and accelerators
- Asymmetric / partitionable memory hierarchies
- Asymmetric / partitionable interconnect
- Asymmetric main memories
Thought Experiment: Asymmetry Everywhere

- **Morph software components** to match asymmetric HW components
  - Multiple versions for different resource characteristics

- **Version 1**: High-power, High perf.
- **Version 2**: Power/performance optimized for each access pattern
- **Version 3**: Different technologies, Power characteristics

- Asymmetric / configurable cores and accelerators
- Asymmetric / partitionable memory hierarchies
- Asymmetric / partitionable interconnect
- Asymmetric main memories
Many Research and Design Questions

- How to design asymmetric components?
  - Fixed, partitionable, reconfigurable components?
  - What types of asymmetry? Access patterns, technologies?

- What monitoring to perform cooperatively in HW/SW?
  - Automatically discover phase/task requirements

- How to design feedback/control loop between components and runtime system software?

- How to design the runtime to automatically manage resources?
  - Track task behavior, pick “best-fit” components for the entire workload
Exploiting Asymmetry: Simple Examples

- Execute critical/serial sections on high-power, high-performance cores/resources [Suleman+ ASPLOS’09, ISCA’10, Top Picks’10’11, Joao+ ASPLOS’12, ISCA’13]
  - Programmer can write less optimized, but more likely correct programs
### Execute each code block on the most efficient execution backend for that block [Fallin+ ICCD’14]
- Enables a much more efficient and still high performance core design

<table>
<thead>
<tr>
<th>High–power</th>
<th>VLIW Backend</th>
</tr>
</thead>
<tbody>
<tr>
<td>High perf.</td>
<td></td>
</tr>
</tbody>
</table>

#### Backend
- Asymmetric / configurable cores and accelerators
- Asymmetric / partitionable memory hierarchies
- Asymmetric / partitionable interconnect
- Asymmetric main memories

<table>
<thead>
<tr>
<th>Power/performance optimized for each access pattern</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Different technologies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Power characteristics</td>
<td></td>
</tr>
</tbody>
</table>

10
Exploiting Asymmetry: Simple Examples

■ Execute streaming “memory phases” on streaming-optimized cores and memory hierarchies
  ■ More efficient and higher performance than general purpose hierarchy
Exploiting Asymmetry: Simple Examples

- Execute bandwidth-sensitive threads on a bandwidth-optimized network, latency-sensitive ones on a latency-optimized network [Das+ DAC’13]
  - Higher performance and energy-efficiency than a single network
Exploiting Asymmetry: Simple Examples

<table>
<thead>
<tr>
<th>High–power High perf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power/performance optimized for each access pattern</td>
</tr>
<tr>
<td>Different technologies Power characteristics</td>
</tr>
</tbody>
</table>

  - Higher performance and energy-efficiency than symmetric/free-for-all

- Higher performance and energy-efficiency than symmetric/free-for-all
Exploiting Asymmetry: Simple Examples

- Have multiple different memory scheduling policies apply them to different sets of threads based on thread behavior [Kim+ MICRO 2010, Top Picks 2011] [Ausavarungnirun+ ISCA 2012]
  - Higher performance and fairness than a homogeneous policy
Exploiting Asymmetry: Simple Examples

- Build main memory with different technologies with different characteristics (e.g., latency, bandwidth, cost, energy, reliability) [Meza+ IEEE CAL’12, Yoon+ ICCD’12, Luo+ DSN’14]
  - Higher performance and energy-efficiency than homogeneous memory
Exploiting Asymmetry: Simple Examples

<table>
<thead>
<tr>
<th>High–power</th>
<th>Low–power</th>
</tr>
</thead>
<tbody>
<tr>
<td>High perf.</td>
<td>Low perf.</td>
</tr>
</tbody>
</table>

- Build main memory with different technologies with different characteristics (e.g., latency, bandwidth, cost, energy, reliability) [Meza+ IEEE CAL’12, Yoon+ ICCD’12, Luo+ DSN’14]
  - Lower-cost than homogeneous-reliability memory at same availability
### Design each memory chip to be heterogeneous to achieve low latency and low energy at reasonably low cost [Lee+ HPCA’13, Liu+ ISCA’12]

- Higher performance and energy-efficiency than single-level memory

#### Table: Heterogeneous Memory Examples

<table>
<thead>
<tr>
<th>High-power</th>
<th>Power/performance optimized for each access pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>High perf.</td>
<td></td>
</tr>
</tbody>
</table>

| Asymmetric / configurable cores and accelerators |
| Asymmetric / partitionable memory hierarchies   |
| Asymmetric / partitionable interconnect         |
| Asymmetric main memories                        |
Some Readings


Multiprocessors
Readings: Multiprocessing

Required

Recommended
Memory Consistency

- **Required**
Readings: Cache Coherence

- **Required**

- **Recommended:**
  - Culler and Singh, *Parallel Computer Architecture*
    - Chapter 5.1 (pp 269 – 283), Chapter 5.3 (pp 291 – 305)
  - P&H, *Computer Organization and Design*
    - Chapter 5.8 (pp 534 – 538 in 4th and 4th revised eds.)
Multiprocessors and Issues in Multiprocessing
Flynn’s Taxonomy of Computers


- **SISD**: Single instruction operates on single data element
- **SIMD**: Single instruction operates on multiple data elements
  - Array processor
  - Vector processor
- **MISD**: Multiple instructions operate on single data element
  - Closest form: systolic array processor, streaming processor
- **MIMD**: Multiple instructions operate on multiple data elements (multiple instruction streams)
  - Multiprocessor
  - Multithreaded processor
SIMD Example: Vector & Array Processors

Array vs. Vector Processors

ARRAY PROCESSOR

VECTOR PROCESSOR

Instruction Stream
LD  VR ← A[3:0]
ADD VR ← VR, 1
MUL VR ← VR, 2
ST  A[3:0] ← VR

Same op @ same time
LD0  LD1  LD2  LD3
AD0  AD1  AD2  AD3
MU0  MU1  MU2  MU3
ST0  ST1  ST2  ST3

Different ops @ time
LD0  LD1  AD0
LD2  AD1  MU0
LD3  AD2  MU1  ST0

Different ops @ same space

Same op @ space
ST0  ST1  ST2  ST3

Digital Design and Comp. Arch. - Lecture 19: SIMD Architectures (Vector and Array Processors)

Onur Mutlu Lectures
37.6K subscribers

3.1K views  Streamed 6 months ago  Livestream - Digital Design and Computer Architecture - ETH Zürich (Spring 2023)
Digital Design and Computer Architecture, ETH Zürich, Spring 2023 https://safari.ethz.ch/digitaltechnik...
An Example Modern Systolic Array: TPU (II)

As reading a large SRAM uses much more power than arithmetic, the matrix unit uses systolic execution to save energy by reducing reads and writes of the Unified Buffer [Kum80][Ram91][Ovt15b]. Figure 4 shows that data flows in from the left, and the weights are loaded from the top. A given 256-element multiply-accumulate operation moves through the matrix as a diagonal wavefront. The weights are preloaded, and take effect with the advancing wave alongside the first data of a new block. Control and data are pipelined to give the illusion that the 256 inputs are read at once, and that they instantly update one location of each of 256 accumulators. From a correctness perspective, software is unaware of the systolic nature of the matrix unit, but for performance, it does worry about the latency of the unit.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.
Why Parallel Computers?

- **Parallelism**: Doing multiple things at a time
- **Things**: instructions, operations, tasks

**Main (or Original) Goal**
- Improve performance (Execution time or task throughput)
  - Execution time of a program governed by Amdahl’s Law

**Other Goals**
- Reduce power consumption
  - (4N units at freq F/4) consume less power than (N units at freq F)
  - Why?
- Improve cost efficiency and scalability, reduce complexity
  - Harder to design a single unit that performs as well as N simpler units
- **Improve dependability**: Redundant execution in space
Types of Parallelism and How to Exploit Them

- **Instruction Level Parallelism**
  - Different instructions within a stream can be executed in parallel
  - Pipelining, out-of-order execution, speculative execution, VLIW
  - Dataflow

- **Data Parallelism**
  - Different pieces of data can be operated on in parallel
  - SIMD: Vector processing, array processing
  - Systolic arrays, streaming processors

- **Task Level Parallelism**
  - Different “tasks/threads” can be executed in parallel
  - Multithreading
  - Multiprocessing (multi-core)
Task-Level Parallelism: Creating Tasks

- Partition a single problem into multiple related tasks (threads)
  - Explicitly: Parallel programming
    - Easy when tasks are natural in the problem
      - Web/database queries
    - Difficult when natural task boundaries are unclear
  - Transparently/implicitly: Thread level speculation
    - Partition a single thread speculatively

- Run many independent tasks (processes) together
  - Easy when there are many processes
    - Batch simulations, different users, cloud computing workloads
  - Does not improve the performance of a single task
Multiprocessing Fundamentals
Multiprocessor Types

- Loosely coupled multiprocessors
  - No shared global memory address space
  - Multicomputer network
    - Network-based multiprocessors
  - Usually programmed via message passing
    - Explicit calls (send, receive) for communication

- Tightly coupled multiprocessors
  - Shared global memory address space
  - Traditional multiprocessing: symmetric multiprocessing (SMP)
    - Existing multi-core processors, multithreaded processors
  - Programming model similar to uniprocessors (i.e., multitasking uniprocessor) except
    - Operations on shared data require synchronization
Main Design Issues in Tightly-Coupled MP

- Shared memory synchronization
  - How to handle synchronization: locks, atomic operations, barriers

- Cache coherence
  - How to ensure correct operation in the presence of private caches keeping the same memory address cached

- Memory consistency: Ordering of all memory operations
  - What should the programmer expect the hardware to provide?

- Shared resource management

- Communication: Interconnects
Main Programming Issues in Tightly-Coupled MP

- **Load imbalance**
  - How to partition a single task into multiple tasks

- **Synchronization**
  - How to synchronize (efficiently) between tasks
  - How to communicate between tasks
  - Locks, barriers, pipeline stages, condition variables, semaphores, atomic operations, ...

- **Contention (avoidance & management)**
- **Maximizing parallelism**
- **Ensuring correct operation while optimizing for performance**
Aside: Hardware-based Multithreading

- Coarse grained
  - Quantum based
  - Event based (switch-on-event multithreading), e.g., switch on L3 miss

- Fine grained
  - Cycle by cycle

- Simultaneous
  - Can dispatch instructions from multiple threads at the same time
  - Good for improving execution unit utilization
Lecture on Fine-Grained Multithreading

**Fine-Grained Multithreading**

- **Idea:** Hardware has multiple thread contexts (PC+registers).
  - Each cycle, fetch engine fetches from a different thread.
  - By the time the fetched branch/instruction resolves, no instruction is fetched from the same thread.
  - Branch/instruction resolution latency overlapped with execution of other threads’ instructions.

- **No logic needed for handling control and data dependences within a thread**
  - Single thread performance suffers
  - Extra logic for keeping thread contexts
  - Does not overlap latency if not enough threads to cover the whole pipeline

https://www.youtube.com/watch?v=6e5KZcCGBYw&list=PL5Q2soXY2Zi_uej3aY39YB5pfW4SJ7LIN&index=16
Coarse-grained Multithreading

- Idea: When a thread is stalled due to some event, switch to a different hardware context
- Switch-on-event multithreading

More on Multithreading (I)

https://www.youtube.com/onurmutlulecutes
More on Multithreading (II)

Carnegie Mellon - Parallel Computer Architecture 2012 - Onur Mutlu - Lecture 10 - Multithreading II

1,594 views • Sep 21, 2013

Carnegie Mellon Computer Architecture
1.8T subscribers

Lecture 10: Multithreading II
Lecturer: Prof. Onur Mutlu (http://users.ece.cmu.edu/~omutlu/)
Date: September 28, 2012.

https://www.youtube.com/onurmutlulectures
More on Multithreading (III)

1,132 views • Sep 21, 2013

Lecture 13: Multi-threading III
Lecturer: Prof. Onur Mutlu (http://users.ece.cmu.edu/~omutlu)
Date: October 5, 2012.

https://www.youtube.com/onurmutlulectures
Lectures on Multithreading

- Parallel Computer Architecture, Fall 2012, Lecture 9
  - Multithreading I (CMU, Fall 2012)
  - https://www.youtube.com/watch?v=iqi9wFqFiNU&list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D&index=51

- Parallel Computer Architecture, Fall 2012, Lecture 10
  - Multithreading II (CMU, Fall 2012)
  - https://www.youtube.com/watch?v=e8lfl6MbILg&list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D&index=52

- Parallel Computer Architecture, Fall 2012, Lecture 13
  - Multithreading III (CMU, Fall 2012)
  - https://www.youtube.com/watch?v=7vkDpZ1-hHM&list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D&index=53

- Parallel Computer Architecture, Fall 2012, Lecture 15
  - Speculation I (CMU, Fall 2012)
  - https://www.youtube.com/watch?v=-hbmzIDe0sA&list=PL5PHm2jkkXmgDN1PLwOY_tGtUlynnyV6D&index=54

https://www.youtube.com/onurmutlulectures
Limits of Parallel Speedup
Parallel Speedup Example

- $a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$

- Assume given inputs: x and each $a_i$

- Assume each operation 1 cycle, no communication cost, each op can be executed in a different processor

- How fast is this with a single processor?
  - Assume no pipelining or concurrent execution of instructions

- How fast is this with 3 processors?
\[ R = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \]

Single processor: 11 operations (data flow graph)

\[ \tau_1 = 11 \text{ cycles} \]
\[ R = a_4 x^4 + a_3 x^3 + a_2 x^2 + a_1 x + a_0 \]

Three processors: \( T_3 \) (execution with 3 proc.)

\[ T_3 = 5 \text{ cycles} \]
Speedup with 3 Processors

\[ T_3 = \frac{5 \text{ cycles}}{} \]

\[ \text{Speedup with 3 processors} = \frac{11}{5} = 2.2 \]

\[ \left( \frac{T_1}{T_3} \right) \]

Is this a fair comparison?
\[ T_1 = 8 \text{ cycles} \]

Speedup with 3 proc. \( = \frac{T_1^{\text{best}}}{T_3^{\text{best}}} = \frac{8}{5} = 1.6 \)

\( (\text{net} 2.2) \)
Superlinear Speedup

- Can speedup be greater than \( P \) with \( P \) processing elements?
- **Unfair comparisons**
  Compare best parallel algorithm to wimpy serial algorithm \( \rightarrow \) unfair
- **Cache/memory effects**
  More processors \( \rightarrow \) more cache or memory \( \rightarrow \) fewer misses in cache/mem
Utilization, Redundancy, Efficiency

- **Traditional metrics**
  - Assume all P processors are tied up for parallel computation

- **Utilization**: How much processing capability is used
  - \( U = \frac{\text{# Operations in parallel version}}{\text{processors} \times \text{Time}} \)

- **Redundancy**: how much extra work is done with parallel processing
  - \( R = \frac{\text{# of operations in parallel version}}{\text{# operations in best single processor algorithm version}} \)

- **Efficiency**
  - \( E = \frac{\text{Time with 1 processor}}{\text{processors} \times \text{Time with P processors}} \)
  - \( E = \frac{U}{R} \)
Utilization of a Multiprocessor

Utilization: How much processing capability we use

$$U = \frac{\text{Ops with } p \text{ proc.}}{p \times T_p}$$

$$U = \frac{10 \text{ operations (in parallel version)}}{3 \text{ processors } \times 5 \text{ time units}} = \frac{10}{15}$$
Redundancy: How much extra work due to multiprocessing

\[ R = \frac{\text{Ops with } p \text{ proc.} \text{ best}}{\text{Ops with } 1 \text{ proc.} \text{ best}} = \frac{10}{8} \]

\( R \) is always \( \geq 1 \)

Efficiency: How much resource we use compared to how much resource we can get away with

\[ E = \frac{1. T_{\text{best}}}{p. T_{\text{best}}} \quad \text{(tying up } p \text{ proc. for } T_p \text{ time units)} \]

\[ = \frac{8}{15} \quad \left( E = \frac{U}{R} \right) \]
Amdahl’s Law and Caveats of Parallelism
Amdahl’s Law

- **Amdahl’s Law**
  - $f$: Parallelizable fraction of a program
  - $N$: Number of processors

\[
\text{Speedup} = \frac{1}{1 - f + \frac{f}{N}}
\]


- **Maximum speedup limited by serial portion: Serial bottleneck**
Caveats of Parallelism (I)

Why the reality? (diminishing returns)

\[ T_p = \alpha \frac{T_1}{p} + (1-\alpha)T_1 \]

- Parallelizable part/ fraction of the single-processor program
- Non-parallelizable part
Amdahl’s Law

\[
\text{Speedup with } p \text{ proc.} = \frac{T_1}{T_p} = \frac{1}{\frac{\alpha}{p} + (1-\alpha)}
\]

\[
\text{Speedup as } p \to \infty = \frac{1}{1-\alpha} \rightarrow \text{bottleneck for parallel Speedup}
\]

Amdahl’s Law Implication 1

Amdahl’s Law illustrated

Adding more and more processors gives less and less benefit, if \( \alpha < 1 \)
Amdahl’s Law Implication 2

The benefit (speedup) is small until $\alpha \approx 1$.
Caveats of Parallelism (II)

- **Amdahl’s Law**
  - \( f \): Parallelizable fraction of a program
  - \( N \): Number of processors

\[
\text{Speedup} = \frac{1}{1 - f + \frac{f}{N}}
\]


- Maximum speedup limited by serial portion: **Serial bottleneck**
- Parallel portion is usually not perfectly parallel
  - **Synchronization** overhead (e.g., updates to shared data)
  - **Load imbalance** overhead (imperfect parallelization)
  - **Resource sharing** overhead (contention among \( N \) processors)
Sequential Bottleneck

![Graph showing the effect of varying N values on speedup as a function of parallel fraction f. The graph compares three cases: N=10, N=100, and N=1000.](image)
Why the Sequential Bottleneck?

- Parallel machines have the sequential bottleneck

- Main cause: Non-parallelizable operations on data (e.g. non-parallelizable loops)
  
  for (i = 0; i < N; i++)
  
  \[ A[i] = (A[i] + A[i-1]) / 2 \]

- There are other causes as well:
  - Single thread prepares data and spawns parallel tasks (usually sequential)
Another Example of Sequential Bottleneck (I)

InitPriorityQueue(PQ);
SpawnThreads();

.ForEachEach Thread:

while (problem not solved)

  Lock (X)
  SubProblem = PQ.remove();
  Unlock(X);

  Solve(SubProblem);
  If(problem solved) break;
  NewSubProblems = Partition(SubProblem);

  Lock(X)
  PQ.insert(NewSubProblems);
  Unlock(X)

  ...

PrintSolution();

LEGEND
A, E: Amdahl’s serial part
B: Parallel Portion
C1, C2: Critical Sections
D: Outside critical section

Another Example of Sequential Bottleneck (II)

Bottlenecks in Parallel Portion

- **Synchronization**: Operations manipulating shared data cannot be parallelized
  - Locks, mutual exclusion, barrier synchronization
  - **Communication**: Tasks may need values from each other
    - Causes thread serialization when shared data is contended

- **Load Imbalance**: Parallel tasks may have different lengths
  - Due to imperfect parallelization or microarchitectural effects
    - Reduces speedup in parallel portion

- **Resource Contention**: Parallel tasks can share hardware resources, delaying each other
  - Replicating all resources (e.g., memory) expensive
    - Additional latency not present when each task runs alone
Bottlenecks in Parallel Portion: Another View

- Threads in a multi-threaded application can be interdependent
  - As opposed to threads from different applications

- Such threads can synchronize with each other
  - Locks, barriers, pipeline stages, condition variables, semaphores, ...

- Some threads can be on the critical path of execution due to synchronization; some threads are not

- Within a thread, some “code segments” may be on the critical path of execution; some are not
Remember: Critical Sections

- Enforce mutually exclusive access to shared data
- Only one thread can be executing it at a time
- Contended critical sections make threads wait → threads causing serialization can be on the critical path

Each thread:

```java
loop {
    Compute
    lock(A)
    Update shared data
    unlock(A)
}
```
Critical Section Example from MySQL

- Critical Section
- Open database tables
- Perform the operations
- Access Open Tables Cache
- Parallel

Graph:
- Chip Area (cores)
- Speedup
- Asymmetric
- Symmetric
Remember: Barriers

- Synchronization point
- Threads have to wait until all threads reach the barrier
- Last thread arriving to the barrier is on the critical path

Each thread:

```java
loop1 {
    Compute
}
barrier
loop2 {
    Compute
}
```
Remember: Stages of Pipelined Programs

- Loop iterations are statically divided into code segments called *stages*
- Threads execute stages on different cores
- Thread executing the slowest stage is on the critical path

```plaintext
loop {
  Compute1
  Compute2
  Compute3
}
```
Difficult in Parallel Programming

- Little difficulty if parallelism is natural
  - “Embarrassingly parallel” applications
  - Multimedia, physical simulation, graphics
  - Large web servers, databases?

- Difficulty is in
  - Getting parallel programs to work correctly
  - Optimizing performance in the presence of bottlenecks

- Much of parallel computer architecture is about
  - Designing machines that overcome the sequential and parallel bottlenecks to achieve higher performance and efficiency
  - Making programmer’s job easier in writing correct and high-performance parallel programs
Some Readings on Bottlenecks & Bottleneck Acceleration
Parallel Application Memory Scheduling

Eiman Ebrahimi, Rustam Miftakhutdinov, Chris Fallin, Chang Joo Lee, Onur Mutlu, and Yale N. Patt,
"Parallel Application Memory Scheduling"
Proceedings of the 44th International Symposium on Microarchitecture (MICRO), Porto Alegre, Brazil, December 2011. Slides (pptx)
Accelerated Critical Sections


One of the 13 computer architecture papers of 2009 selected as Top Picks by IEEE Micro.

Accelerating Critical Section Execution with Asymmetric Multi-Core Architectures

M. Aater Suleman
University of Texas at Austin
suleman@hps.utexas.edu

Onur Mutlu
Carnegie Mellon University
onur@cmu.edu

Moinuddin K. Qureshi
IBM Research
mkquresh@us.ibm.com

Yale N. Patt
University of Texas at Austin
patt@ece.utexas.edu
Bottleneck Identification & Scheduling

Jose A. Joao, M. Aater Suleman, Onur Mutlu, and Yale N. Patt, "Bottleneck Identification and Scheduling in Multithreaded Applications"

Bottleneck Identification and Scheduling in Multithreaded Applications

José A. Joao
ECE Department
The University of Texas at Austin
joao@ece.utexas.edu

M. Aater Suleman
Calxeda Inc.
aater.suleman@calxeda.com

Onur Mutlu
Computer Architecture Lab.
Carnegie Mellon University
onur@cmu.edu

Yale N. Patt
ECE Department
The University of Texas at Austin
patt@ece.utexas.edu
Utility-Based Acceleration


Utility-Based Acceleration of Multithreaded Applications on Asymmetric CMPs

José A. Joao † M. Aater Suleman ‡‡ Onur Mutlu § Yale N. Patt †

† ECE Department
The University of Texas at Austin
Austin, TX, USA
{joao, patt}@ece.utexas.edu

‡‡ Flux7 Consulting
Austin, TX, USA
suleman@hps.utexas.edu

§ Computer Architecture Laboratory
Carnegie Mellon University
Pittsburgh, PA, USA
onur@cmu.edu
Data Marshaling

- M. Aater Suleman, Onur Mutlu, Jose A. Joao, Khubaib, and Yale N. Patt, "Data Marshaling for Multi-core Architectures"

One of the 11 computer architecture papers of 2010 selected as Top Picks by IEEE Micro.

Data Marshaling for Multi-core Architectures

M. Aater Suleman† Onur Mutlu§ José A. Joao† Khubaib† Yale N. Patt†

†The University of Texas at Austin
{suleman, joao, khubaib, patt}@hps.utexas.edu

§Carnegie Mellon University
onur@cmu.edu
Lectures on Bottleneck Acceleration

- Lecture 18: Parallelism and Heterogeneity
  - Comp Arch, ETH Zurich, Fall 2023
  - https://www.youtube.com/live/r1GmCoVQ_hA?si=5t0kRN2l5sG8YhSa

- Lecture 17a: Parallelism and Heterogeneity
  - Comp Arch, ETH Zurich, Fall 2021
  - https://www.youtube.com/watch?v=GLzG_rEDn9A&list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF&index=18

- Lecture 18a: Bottleneck Acceleration
  - Comp Arch, ETH Zurich, Fall 2021
  - https://www.youtube.com/watch?v=P8l3SMAbYw&list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF&index=19
Lecture on Parallelism & Heterogeneity

ACMP Performance vs. Parallelism

Area-budget = 16 small cores

<table>
<thead>
<tr>
<th></th>
<th>Large Core</th>
<th>Large Core</th>
<th>Small Core</th>
<th>Small Core</th>
<th>Small Core</th>
<th>Small Core</th>
<th>Small Core</th>
<th>Small Core</th>
<th>Large Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Tile-Large&quot;</td>
<td>4</td>
<td>0</td>
<td>16</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Tile-Small&quot;</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACMP</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parallel Throughput

- 2 x 4 = 8
- 1 x 16 = 16
- 1 x 2 + 1 x 12 = 14

https://www.youtube.com/watch?v=GLzG_rEDn9A&list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF&index=19
Lecture on Bottleneck Acceleration

Bottleneck Acceleration

- Small Core 1
  - Acceleration Index Table (AIT)
    - bid=x4700, large core 0

- Small Core
  - AIT
    - bid=x4700, large core 0

- Large Core 0
  - Scheduling Buffer (SB)
    - bid=x4600, twc=100
    - bid=x4700, twc=10000

Bottleneck Table (BT)

https://www.youtube.com/watch?v=P8l3SMAbyYw&list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF&index=19
Computer Architecture
Lecture 19a: Multiprocessors

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich
Fall 2023
30 November 2023
An Example Parallel Problem: Task Assignment to Processors
Static versus Dynamic Scheduling

- **Static**: Done at compile time or parallel task creation time
  - Schedule does not change based on runtime information

- **Dynamic**: Done at run time (e.g., after tasks are created)
  - Schedule changes based on runtime information

- **Example**: Instruction scheduling
  - Why would you like to do dynamic scheduling?
  - What pieces of information are not available to the static scheduler?
Parallel Task Assignment: Tradeoffs

- Problem: N tasks, P processors, N>P. Do we assign tasks to processors statically (fixed) or dynamically (adaptive)?

- Static assignment
  + Simpler: No movement of tasks.
  - Inefficient: Underutilizes resources when load is not balanced
    
    *When can load not be balanced?*

- Dynamic assignment
  + Efficient: Better utilizes processors when load is not balanced
  - More complex: Need to move tasks to balance processor load
  - Higher overhead: Task movement takes time, can disrupt locality
Parallel Task Assignment: Example

- Compute histogram of a large set of values
- Parallelization:
  - Divide the values across $T$ tasks
  - Each task computes a local histogram for its value set
  - Local histograms merged with global histograms in the end
Parallel Task Assignment: Example (II)

- How to schedule tasks updating local histograms?
  - Static: Assign equal number of tasks to each processor
  - Dynamic: Assign tasks to a processor that is available
  - When does static work as well as dynamic?

- Implementation of Dynamic Assignment with Task Queues

![Diagram](image)

(a) Distributed Task Stealing  
(b) Hierarchical Task Queuing
Software Task Queues

- What are the advantages and disadvantages of each?
  - Centralized
  - Distributed
  - Hierarchical

(a) Distributed Task Stealing
(b) Hierarchical Task Queuing
Task Stealing

- **Idea:** When a processor’s task queue is empty it steals a task from another processor’s task queue
  - Whom to steal from? (Randomized stealing works well)
  - How many tasks to steal?

+ Dynamic balancing of computation load

- Additional communication/synchronization overhead between processors
- Need to stop stealing if no tasks to steal
Parallel Task Assignment: Tradeoffs

- Who does the assignment? Hardware versus software?

  - Software
    - + Better scope
    - - More time overhead
    - - Slow to adapt to dynamic events (e.g., a processor becoming idle)

  - Hardware
    - + Low time overhead
    - + Can adjust to dynamic events faster
    - - Requires hardware changes (area and possibly energy overhead)
How Can the Hardware Help?

- Managing task queues in software has overhead
  - Especially high when task sizes are small

- An idea: Hardware Task Queues
  - Each processor has a dedicated task queue
  - Software fills the task queues (on demand)
  - Hardware manages movement of tasks from queue to queue
  - There can be a global task queue as well → hierarchical tasking in hardware

  - Optional reading
Dynamic Task Generation

- Does static task assignment work in this case?

- Problem: Searching the exit of a maze

```python
while (problem not solved):
    SubProblem = PriorityQ.remove()
    Solve(SubProblem)
    if (solved):
        break
    NewSubProblems = Partition(SubProblem)
    PriorityQ.insert(NewSubProblems)
```
Programming Model vs. Hardware Execution Model
Programming Models vs. Architectures

- Five major models
  - (Sequential)
  - Shared memory
  - Message passing
  - Data parallel (SIMD)
  - Dataflow
  - Systolic

- Hybrid models?
Shared Memory vs. Message Passing

- Are these programming models or execution models supported by the hardware architecture?

- Does a multiprocessor that is programmed by “shared memory programming model” have to support a shared address space processors?

- Does a multiprocessor that is programmed by “message passing programming model” have to have no shared address space between processors?
Programming Models: Message Passing vs. Shared Memory

- Difference: how communication is achieved between tasks
- **Message passing programming model**
  - Explicit communication via messages
  - Loose coupling of program components
  - Analogy: telephone call or letter, no shared location accessible to all
- **Shared memory programming model**
  - Implicit communication via memory operations (load/store)
  - Tight coupling of program components
  - Analogy: bulletin board, post information at a shared space

Suitability of the programming model depends on the problem to be solved. Issues affected by the model include:
- Overhead, scalability, ease of programming, bugs, match to underlying hardware, ...
Message Passing vs. Shared Memory Hardware

- **Difference:** how task communication is supported in hardware

**Shared memory hardware (or machine model)**

- All processors see a global shared address space
  - Ability to access all memory from each processor
- A write to a location is visible to the reads of other processors

**Message passing hardware (machine model)**

- No global shared address space
- Send and receive variants are the only method of communication between processors (much like networks of workstations today, i.e. clusters)

- Suitability of the hardware depends on the problem to be solved as well as the programming model.
Most of parallel computing history, there was no separation between programming model and hardware

- Message passing: Caltech Cosmic Cube, Intel Hypercube, Intel Paragon
- Shared memory: CMU C.mmp, Sequent Balance, SGI Origin.
- SIMD: ILLIAC IV, CM-1

However, any hardware can really support any programming model

Why?

- Application $\rightarrow$ compiler/library $\rightarrow$ OS services $\rightarrow$ hardware