

Languages and Abstractions for High-Performance
Scientific Computing
CS598 APK

Andreas Kloeckner

Spring 2025

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Why this class?

- ▶ Setting: Performance-Constrained Code
When is a code performance-constrained?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the first question.

- ▶ If your code is performance-constrained, what is the *best* approach?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the second question.

- ▶ If your code is performance-constrained, what is the *second-best* approach?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the third question.

Examples of Performance-Constrained Codes



Discussion:

- ▶ In what way are these codes constrained?
- ▶ How do these scale in terms of the problem size?

What Problem are we Trying To Solve?

$$(C_{ij})_{i,j=1}^{m,n} = \sum_{k=1}^{\ell} A_{ik} B_{kj}$$

- ▶ [Reference BLAS DGEMM code](#)
- ▶ [OpenBLAS DGEMM code](#)

[Demo: intro/DGEMM Performance](#)

Goals: What are we Allowed to Ask For?

- ▶ Goal: “make efficient use of the machine”
- ▶ In general: not an easy question to answer
- ▶ In theory: limited by *some* peak machine throughput
 - ▶ Memory Access
 - ▶ Compute
- ▶ In practice: many other limits (Instruction cache, TLB, memory hierarchy, NUMA, registers)

Class web page

<https://bit.ly/hpcabstr-s25>

contains:

- ▶ Class outline
- ▶ Slides/demos/materials
- ▶ Assignments
- ▶ Discussion Forum
- ▶ Grading Policies
- ▶ Video
- ▶ HW1 (soon)

Welcome Survey

Please go to:

<https://bit.ly/hpcabstr-s25>

and click on 'Start Activity'.

If you are seeing this later, you can find the activity at [Activity: welcome-survey](#).

Grading / Workload

Four components:

- ▶ Homework: 25%
- ▶ Paper Presentation: 25%
 - ▶ 30 minutes (two per class)
 - ▶ Presentation sessions scheduled throughout the semester
 - ▶ Paper list on web page
 - ▶ Sign-up survey: soon
- ▶ Paper Reactions: 10%
- ▶ Computational Project: 40%

Open Source <3

These notes (and the accompanying demos) are open-source!

Bug reports and pull requests welcome:

<https://github.com/inducer/hpc-lang-abstr-notes>

Copyright (C) 2010-2013 Andreas Kloeckner

Copyright (C) 2025 University of Illinois Board of Trustees

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Approaches to High Performance

- ▶ Libraries (seen)
- ▶ Black-box Optimizing Compilers
- ▶ Compilers with Directives
- ▶ Code Transform Systems
- ▶ “Active Libraries”

Q: Give examples of the latter two.



Libraries: A Case Study

$$(C_{ij})_{i,j=1}^{m,n} = \sum_{k=1}^{\ell} A_{ik} B_{kj}$$

Demo: intro/DGEMM Performance



Do Libraries Stand a Chance? (in general)

- ▶ Tremendously successful approach — Name some examples



- ▶ Saw: Three simple integer parameters suffice to lose 'good' performance
 - ▶ Recent efforts: e.g. [Batch BLAS](#)
- ▶ Separation of Concerns

Example: Finite differences – e.g. implement ∂_x , ∂_y , ∂_z as separate (library) subroutines — What is the problem?



- ▶ Flexibility and composition

(Black-Box) Optimizing Compiler: Challenges

Why is black-box optimizing compilation so difficult?

- ▶ Application developer knowledge lost
 - ▶ Simple example: “Rough” matrix sizes
 - ▶ Data-dependent control flow
 - ▶ Data-dependent access patterns
 - ▶ Activities of other, possibly concurrent parts of the program
 - ▶ Profile-guided optimization can recover some knowledge
- ▶ Obtain proofs of required properties
- ▶ Size of the search space

Consider <http://polaris.cs.uiuc.edu/publications/padua.pdf>

Directive-Based Compiler: Challenges

What is a directive-based compiler?

- ▶ Generally same as optimizing compiler
- ▶ Make use of extra promises made by the user
- ▶ What should the user promise?
- ▶ Ideally: feedback cycle between compiler and user
 - ▶ Often broken in both directions
 - ▶ User may not know what the compiler did
 - ▶ Compiler may not be able to express what it needs
- ▶ Directives: generally not mandatory

Lies, Lies Everywhere

- ▶ Semantics form a contract between programmer and language/environment
- ▶ Within those bounds, implementation has full freedom
- ▶ True at every level:
 - ▶ Assembly
 - ▶ “High-level” language (C)

Give examples of lies at these levels:



One approach: *Lie to yourself*

- ▶ “Domain-specific languages” ← A fresh language, I can do what I want!
- ▶ Consistent semantics are notoriously hard to develop
 - ▶ Especially as soon as you start allowing subsets of even (e.g.) C’s

Class Outline

High-level Sections:

- ▶ Intro, Armchair-level Computer Architecture
- ▶ Machine Abstractions
- ▶ Performance: Expectation, Experiment, Observation
- ▶ Programming Languages for Performance
- ▶ Program Representation and Optimization Strategies
- ▶ Code Generation/JIT

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Moore's Law

Moore's Law: The number of transistors on microchips has doubled every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World
in Data

Transistor count

50,000,000,000

10,000,000,000

5,000,000,000

1,000,000,000

500,000,000

100,000,000

50,000,000

10,000,000

5,000,000

1,000,000

500,000

100,000

50,000

10,000

5,000

1,000

500

250

125

62.5

31.25

15.625

7.8125

3.90625

1.953125

0.9765625

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

OurWorldinData.org – Research and data to make progress against the world's largest problem

Issue: More transistors =
faster?

$$\frac{\text{Work}}{s} = \text{Clock Frequency} \times \text{Work/Clock}$$

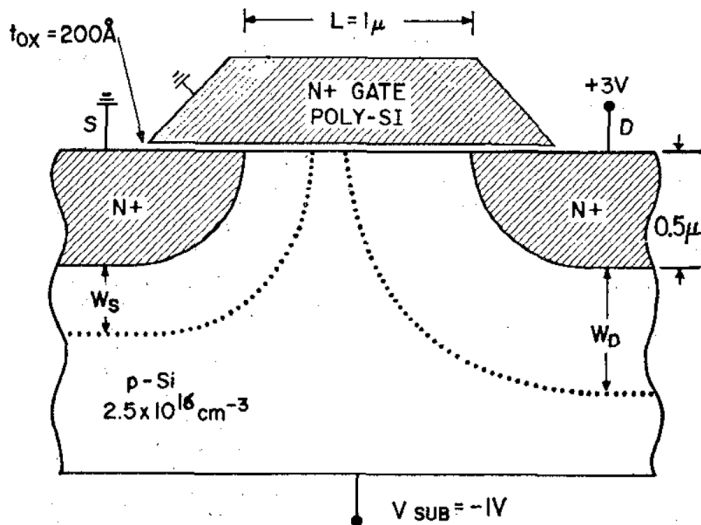
Dennard Scaling of MOSFETs

Parameter	Factor
Dimension	$1/\kappa$
Voltage	$1/\kappa$
Current	$1/\kappa$
Capacitance	$1/\kappa$
Delay Time	$1/\kappa$
Power dissipation/circuit	$1/\kappa^2$
Power density	1

[Dennard et al. '74, via Bohr '07]

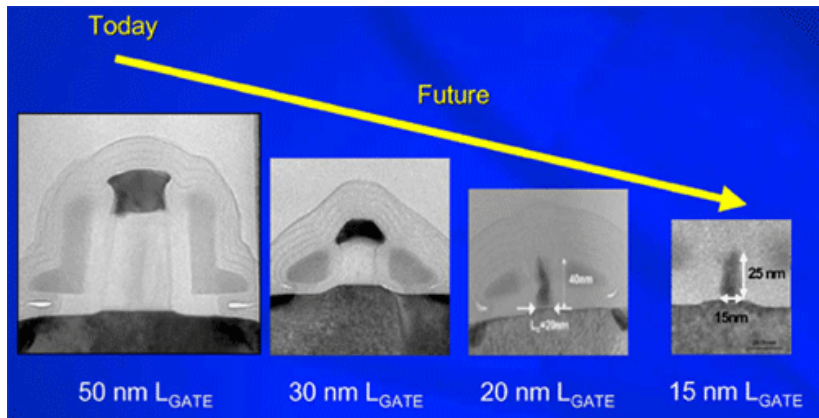
- Frequency = Delay time⁻¹

MOSFETs (“CMOS” – “complementary” MOS): Schematic



[Dennard et al. '74]

MOSFETs: Scaling



[Intel Corp.]

- ▶ 'New' problem at small scale:
Sub-threshold leakage (due to low voltage, small structure)
Dennard scaling is over – and has been for a while.

Peak Architectural Instructions per Clock: Intel

CPU	IPC	Year
Pentium 1	1.1	1993
Pentium 3	1.9	1999
Pentium 4 (Gallatin)	1.9	20
Pentium D	2	2005
Pentium M	2.5	2003
Core 2	3	2006
Sandy Bridge...	<4	2011
Skylake	<4	2015
Golden Cove	<6	2021
Lion Cove	<8	2024

[Charlie Brej <http://brej.org/blog/?p=15>, Wikipedia, Intel]

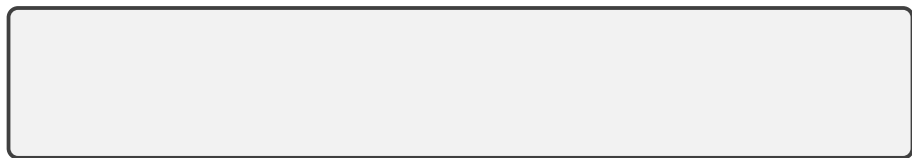
Context: [Lemire: simdjson achieved IPC, '19](#)

Discuss: How do we get out of this dilemma?

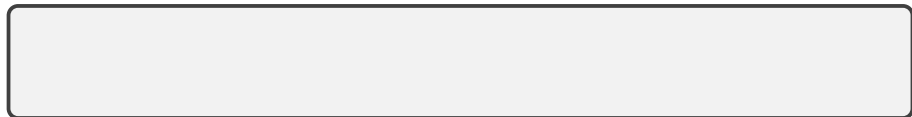
The Performance Dilemma

- ▶ IPC: Brick-ish Wall
- ▶ Clock Frequency: Brick Wall

Ideas:

A large, empty rectangular box with a thin black border and rounded corners, intended for writing ideas.

Question: What is the *conceptual* difference between those ideas?

A large, empty rectangular box with a thin black border and rounded corners, intended for writing the answer to the question.

The Performance Dilemma: Another Look

- ▶ **Really:** A crisis of the 'starts-at-the-top-ends-at-the-bottom' programming model
- ▶ **Tough luck:** Most of our codes are written that way
- ▶ **Even tougher luck:** Everybody on the planet is *trained* to write codes this way

So:

- ▶ **Need:** Different tools/abstractions to write those codes

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

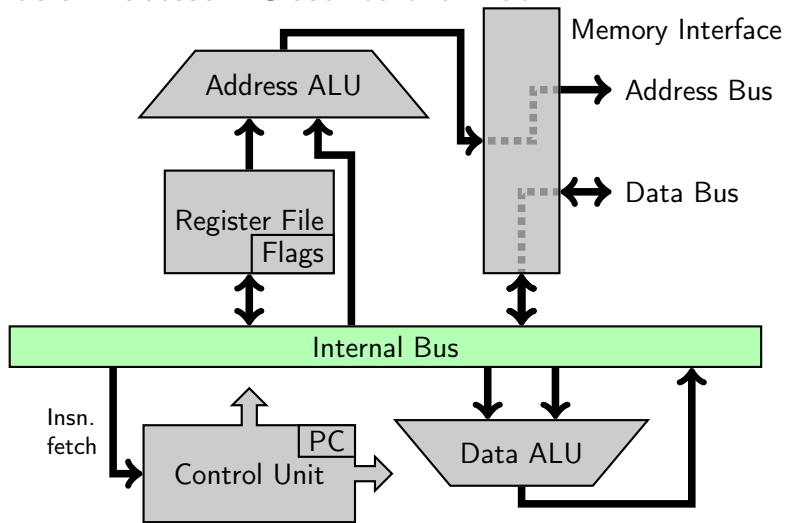
Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

A Basic Processor: Closer to the Truth



- ▶ loosely based on Intel 8086
- ▶ What's a bus?

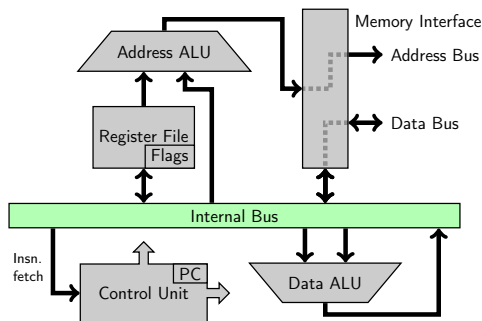
A Very Simple Program

	4:	c7 45 f4 05 00 00 00	movl	\$0x5, -0xc(%rbp)
int a = 5;	b:	c7 45 f8 11 00 00 00	movl	\$0x11, -0x8(%rbp)
int b = 17;	12:	8b 45 f4	mov	-0xc(%rbp), %eax
int z = a * b;	15:	0f af 45 f8	imul	-0x8(%rbp), %eax
	19:	89 45 fc	mov	%eax, -0x4(%rbp)
	1c:	8b 45 fc	mov	-0x4(%rbp), %eax

Things to know:

- ▶ Question: Which is it?
 - ▶ <opcode> <src>, <dest>
 - ▶ <opcode> <dest>, <src>
- ▶ [Addressing modes](#) (Immediate, Register, Base plus Offset)
- ▶ [0xHexadecimal](#)

A Very Simple Program: Another Look



```
4:      c7 45 f4 05 00 00 00 movl    $0x5,-0xc(%rbp)
b:      c7 45 f8 11 00 00 00 movl    $0x11,-0x8(%rbp)
12:     8b 45 f4                mov     -0xc(%rbp),%eax
15:     0f af 45 f8            imul    -0x8(%rbp),%eax
19:     89 45 fc                mov     %eax,-0x4(%rbp)
1c:     8b 45 fc                mov     -0x4(%rbp),%eax
```

A Very Simple Program: Intel Form

```
4:      c7 45 f4 05 00 00 00      mov     DWORD PTR [rbp-0xc],0x5
b:      c7 45 f8 11 00 00 00      mov     DWORD PTR [rbp-0x8],0x11
12:     8b 45 f4                  mov     eax,DWORD PTR [rbp-0xc]
15:     0f af 45 f8              imul    eax,DWORD PTR [rbp-0x8]
19:     89 45 fc                  mov     DWORD PTR [rbp-0x4],eax
1c:     8b 45 fc                  mov     eax,DWORD PTR [rbp-0x4]
```

- ▶ “Intel Form”: (you might see this on the net)
 <opcode> <sized dest>, <sized source>
- ▶ Previous: “AT&T Form”: (we’ll use this)
- ▶ Goal: Reading comprehension.
- ▶ Don’t understand an opcode?

https://en.wikipedia.org/wiki/X86_instruction_listings

Assembly Loops

<code>int main()</code>	0: 55	<code>push %rbp</code>
<code>{</code>	1: 48 89 e5	<code>mov %rsp,%rbp</code>
<code>int y = 0, i;</code>	4: c7 45 f8 00 00 00 00	<code>movl \$0x0,-0x8(%rbp)</code>
<code>for (i = 0;</code>	b: c7 45 fc 00 00 00 00	<code>movl \$0x0,-0x4(%rbp)</code>
<code>y < 10; ++i)</code>	12: eb 0a	<code>jmp 1e <main+0x1e></code>
<code>y += i;</code>	14: 8b 45 fc	<code>mov -0x4(%rbp),%eax</code>
<code>return y;</code>	17: 01 45 f8	<code>add %eax,-0x8(%rbp)</code>
<code>}</code>	1a: 83 45 fc 01	<code>addl \$0x1,-0x4(%rbp)</code>
	1e: 83 7d f8 09	<code>cmpl \$0x9,-0x8(%rbp)</code>
	22: 7e f0	<code>jle 14 <main+0x14></code>
	24: 8b 45 f8	<code>mov -0x8(%rbp),%eax</code>
	27: c9	<code>leaveq</code>
	28: c3	<code>retq</code>

Things to know:

- ▶ [Condition Codes \(Flags\)](#): Zero, Sign, Carry, etc.
- ▶ [Call Stack](#): Stack frame, stack pointer, base pointer
- ▶ [ABI](#): Calling conventions

Demos

Demo: intro/Assembly Reading Comprehension

Demo: Source-to-assembly mapping

Code to try:

```
int main()  
{  
    int y = 0, i;  
    for (i = 0; y < 100; ++i)  
        y += i*i;  
    return y;  
}
```

Also try <https://godbolt.org> for direct source-to-assembly mapping

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Modern Processors?

All of this can be built in about 4000 transistors.

(e.g. MOS 6502 in Apple II, Commodore 64, Atari 2600)

So what exactly are Intel/ARM/AMD/Nvidia doing with the other **billions** of transistors?

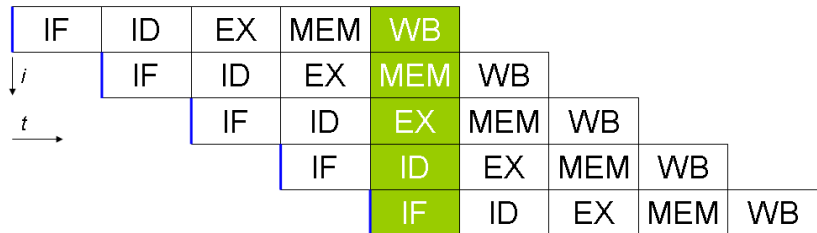
Execution in a Simple Processor



- ▶ [IF] Instruction fetch
- ▶ [ID] Instruction Decode
- ▶ [EX] Execution
- ▶ [MEM] Memory Read/Write
- ▶ [WB] Result Writeback

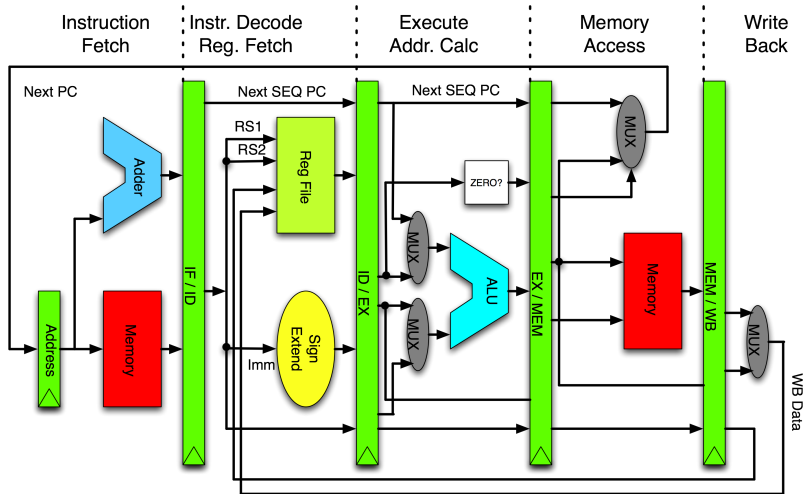
[Wikipedia ©]

Solution: Pipelining



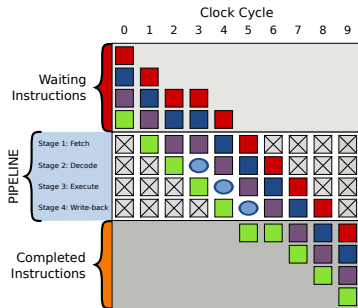
[Wikipedia ©]

MIPS Pipeline: 110,000 transistors




[Wikipedia ©]

Hazards and Bubbles



Q: Types of Pipeline Hazards? (aka: what can go wrong?)

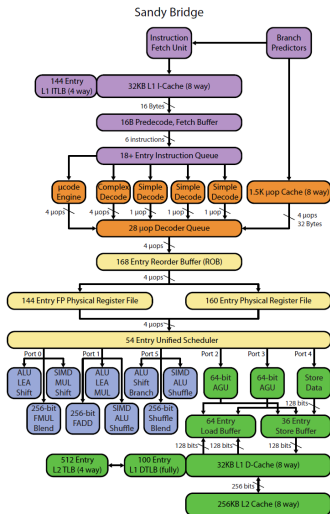
[Wikipedia 

Demo

[Demo: intro/Pipeline Performance Mysteries](#)



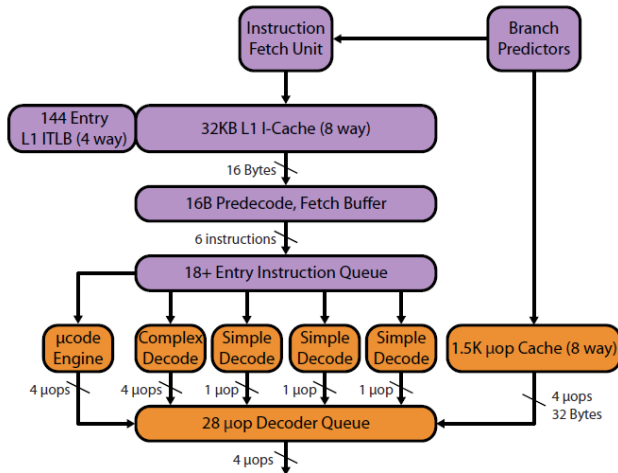
A Glimpse of a More Modern Processor



[David Kanter / Realworldtech.com]

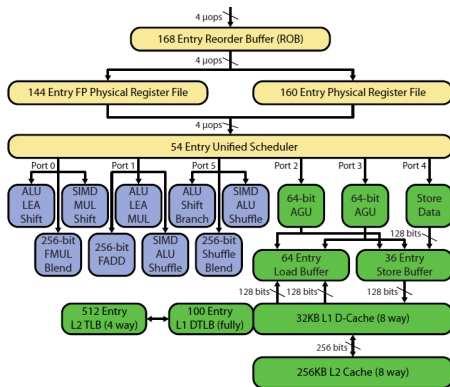
A Glimpse of a More Modern Processor: Frontend

Sandy Bridge



[David Kanter / Realworldtech.com]

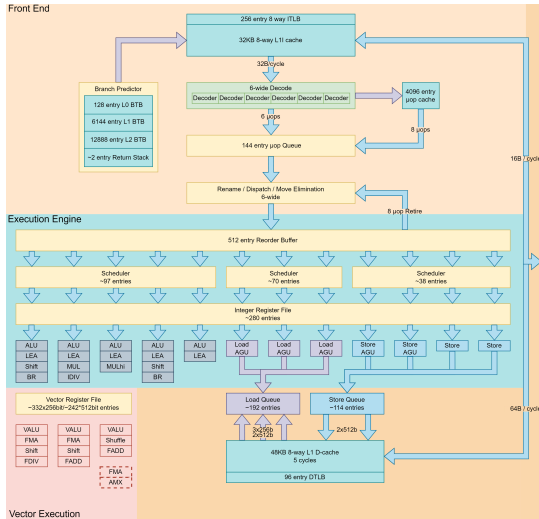
A Glimpse of a More Modern Processor: Backend



- ▶ **New concept:** Instruction-level parallelism (“ILP”, “superscalar”)
- ▶ Where does the IPC number from earlier come from?

[David Kanter / Realworldtech.com]

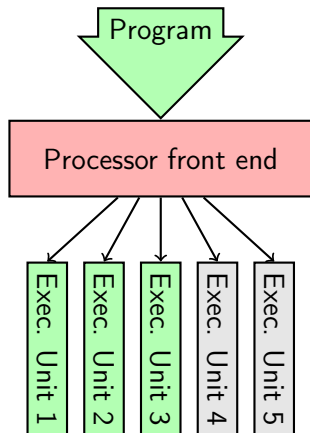
A Glimpse of a More Modern Processor: Golden Cove



Demo

[Demo: intro/More Pipeline Mysteries](#)

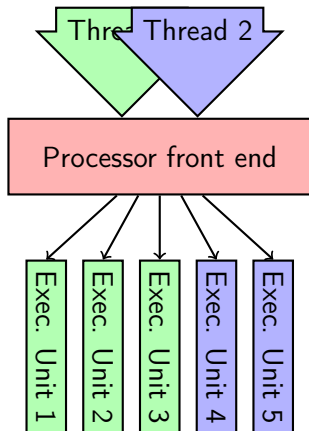
SMT/“Hyperthreading”



Q: Potential issues?



SMT/“Hyperthreading”



Q: Potential issues?



Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

More Bad News from Dennard

Parameter	Factor
Dimension	$1/\kappa$
Line Resistance	κ
Voltage drop	κ
Response time	1
Current density	κ

[Dennard et al. '74, via Bohr '07]

- ▶ The above scaling law is for on-chip interconnects.
- ▶ Current \sim Power vs. response time

Getting information from

- ▶ processor to memory
- ▶ one computer to the next

is

- ▶ slow (in *latency*)
- ▶ power-hungry

Somewhere Behind the Interconnect: Memory

Performance characteristics of memory:

- ▶ Bandwidth
- ▶ Latency

Flops are cheap

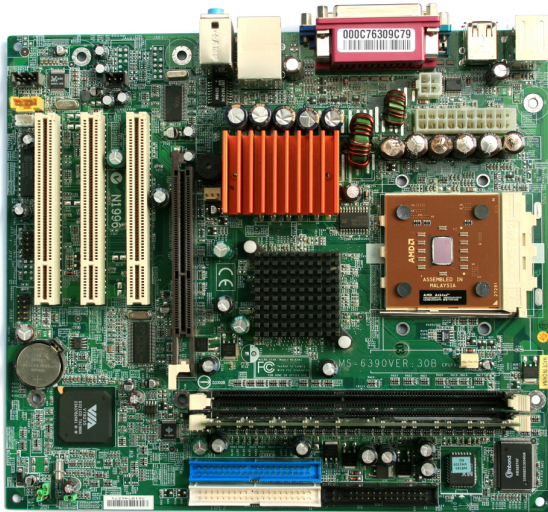
Bandwidth is money

Latency is physics — M. Hoemmen

Minor addition (but important for us)?

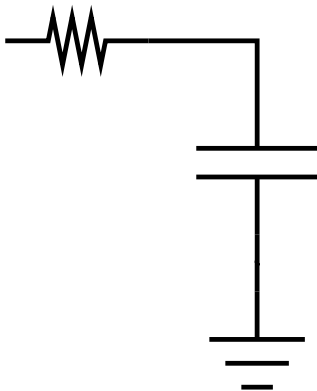


Latency is Physics: Distance

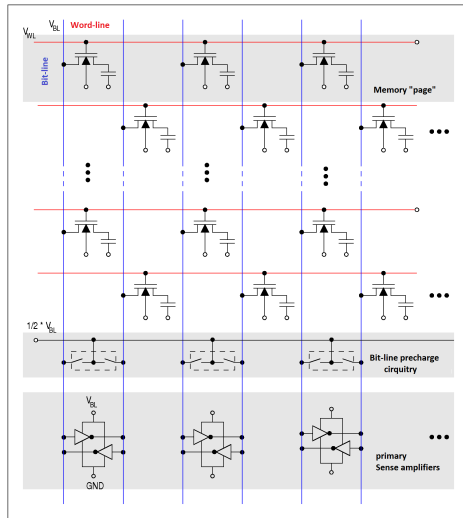


[Wikipedia ©]

Latency is Physics: Electrical Model



Latency is Physics: DRAM



[Wikipedia (cc)]

Latency is Physics: Performance Impact?

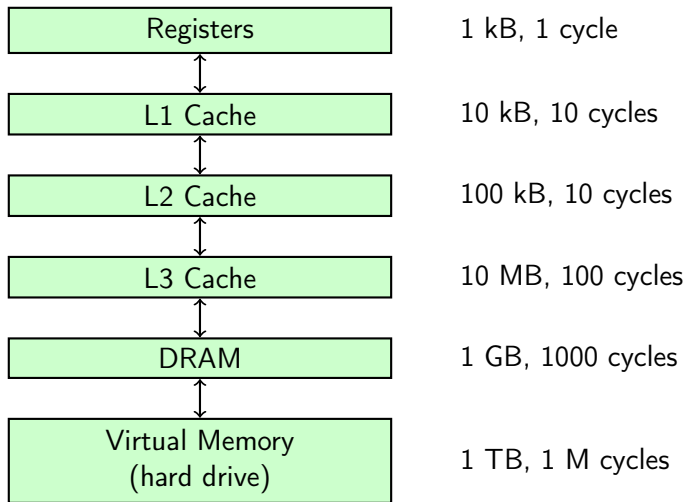
What is the performance impact of high memory latency?



Idea:

- ▶ Put a look-up table of recently-used data onto the chip.
- ▶ Cache

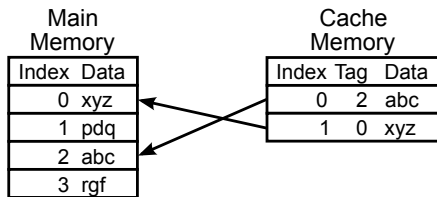
Memory Hierarchy



A Basic Cache

Demands on cache implementation:

- ▶ Fast, small, cheap, low power
- ▶ Fine-grained
- ▶ High “hit”-rate (few “misses”)



Design Goals: at odds with each other. Why?

[Wikipedia 

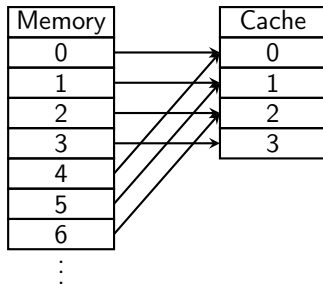
Caches: Engineering Trade-Offs

Engineering Decisions:

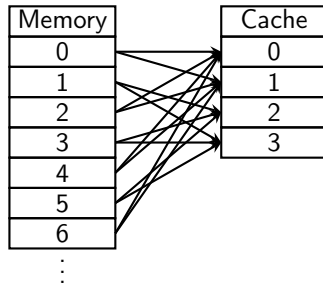
- ▶ More data per unit of access matching logic
→ Larger “Cache Lines”
- ▶ Simpler/less access matching logic
→ Less than full “Associativity”
- ▶ Eviction strategy
- ▶ Size

Associativity

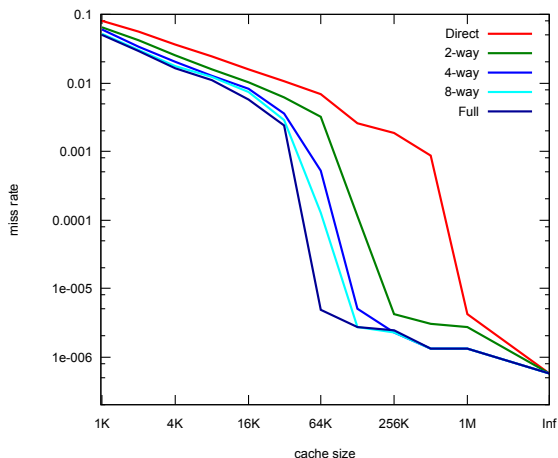
Direct Mapped:



2-way set associative:



Size/Associativity vs Hit Rate



Miss rate versus cache size on the Integer portion of SPEC CPU2000
[Cantin, Hill 2003]

Demo: Learning about Caches

[Demo: intro/Cache Organization on Your Machine](#)

Experiments: 1. Strides: Setup

```
int go(unsigned count, unsigned stride)
{
    const unsigned array_size = 64 * 1024 * 1024;
    int *ary = (int *) malloc(sizeof(int) * array_size);

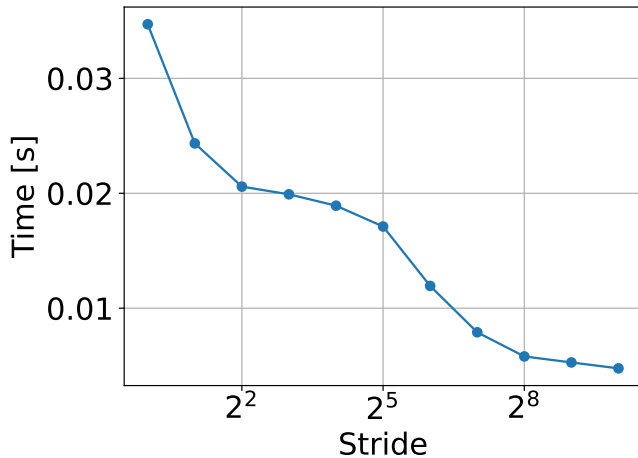
    for (unsigned it = 0; it < count; ++it)
    {
        for (unsigned i = 0; i < array_size; i += stride)
            ary[i] *= 17;
    }

    int result = 0;
    for (unsigned i = 0; i < array_size; ++i)
        result += ary[i];

    free(ary);
    return result;
}
```

What do you expect? [[Ostrovsky '10](#)]

Experiments: 1. Strides: Results



Experiments: 2. Bandwidth: Setup

```
int go(unsigned array_size, unsigned steps)
{
    int *ary = (int *) malloc(sizeof(int) * array_size);
    unsigned asm1 = array_size - 1;

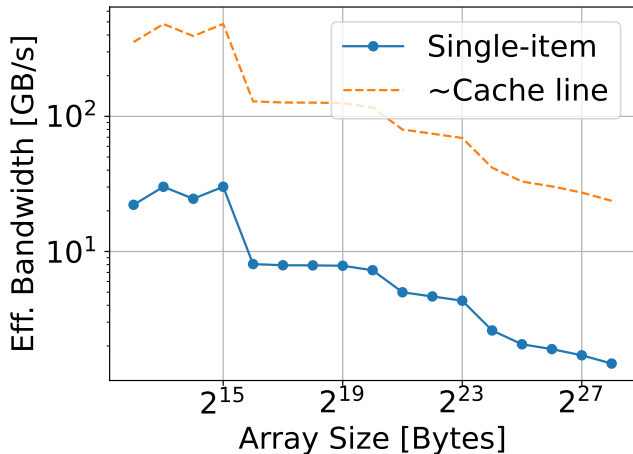
    for (unsigned i = 0; i < 100*steps;)
    {
        #define ONE ary[(i++*16) & asm1] ++;
        #define FIVE ONE ONE ONE ONE ONE ONE
        #define TEN FIVE FIVE
        #define FIFTY TEN TEN TEN TEN TEN
        #define HUNDRED FIFTY FIFTY
        HUNDRED
    }

    int result = 0;
    for (unsigned i = 0; i < array_size; ++i)
        result += ary[i];

    free(ary);
    return result;
}
```

What do you expect? [Ostrovsky '10]

Experiments: 2. Bandwidth: Results



Experiments: 3. A Mystery: Setup

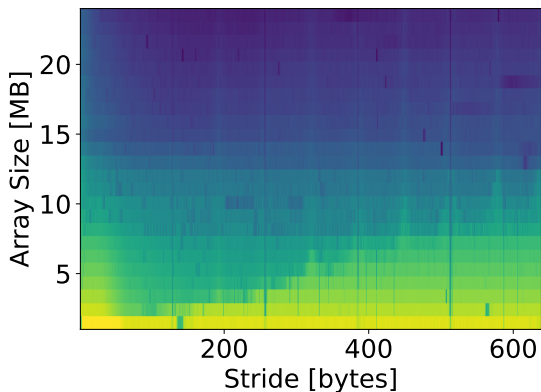
```
int go(unsigned array_size, unsigned stride, unsigned steps)
{
    char *ary = (char *) malloc(sizeof(int) * array_size);

    unsigned p = 0;
    for (unsigned i = 0; i < steps; ++i)
    {
        ary[p] ++;
        p += stride;
        if (p >= array_size)
            p = 0;
    }

    int result = 0;
    for (unsigned i = 0; i < array_size; ++i)
        result += ary[i];

    free(ary);
    return result;
}
```

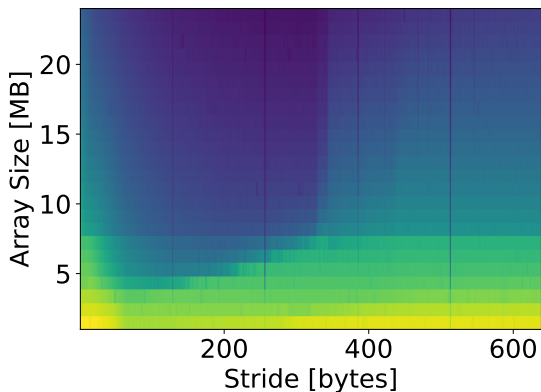

Experiments: 3. A Mystery: Results



Color represents achieved bandwidth:

- Yellow: high
- Blue: low

Experiments: 3. A Mystery: Results on Sandy Bridge



Color represents achieved bandwidth:

- Yellow: high
- Blue: low

Thinking about the Memory Hierarchy

- ▶ What is a **working set**?
- ▶ What is **data locality** of an algorithm?
- ▶ What does this have to do with caches?

Case Study: Streaming Workloads

Q: Estimate expected throughput for saxpy on an architecture with caches. What are the right units?

$$z_i = \alpha x_i + y_i \quad (i = 1, \dots, n)$$



Demo: https://github.com/lcw/stream_ispc

Special Store Instructions

At least two aspects to keep apart:



What hardware behavior might result from these aspects?

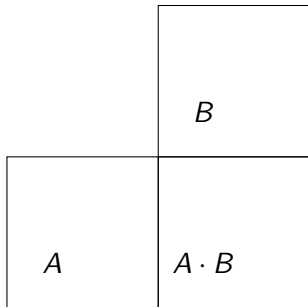


- ▶ Comment on what a compiler can promise on these aspects.
- ▶ Might these 'flags' apply to loads/prefetches?

(see also: [[McCalpin '18](#)])

Case study: Matrix-Matrix Mult. ('MMM'): Code Structure

- ▶ How would you structure a high-performance MMM?
- ▶ What are sources of concurrency?
- ▶ What should you consider your working set?



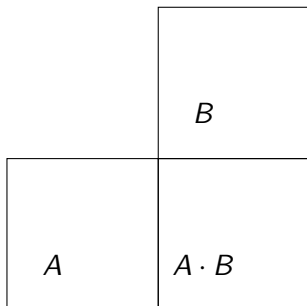
Case study: Matrix-Matrix Mult. ('MMM'): Code Structure

```
for (i)
  for (j)
    for (k)
      c[i,j] += a[i,k]*b[k,j]
```

After:

```
for (itile)
  for (jtile)
    for (ktile)
      for (iinner)
        for (jinner)
          for (kinner)
            c[itile*NB+iinner,...]
              += ...
```

How many tiles?



Case study: Matrix-Matrix Mult. ('MMM') via Latency

Cost model for MMM in a two-level hierarchy based on latency?



[[Yotov et al. '07](#)]

Case study: Matrix-Matrix Mult. ('MMM') via Bandwidth

Cost model for MMM in a two-level hierarchy based on bandwidth?



Case study: Matrix-Matrix Mult. ('MMM'): Discussion

Discussion: What are the main simplifications in each model?



[[Yotov et al. '07](#)]

Q: How can we analyze cache cost of algorithms in general?

Hong/Kung: Red/Blue Pebble Game

Simple means of I/O cost analysis: “Red/blue pebble game”

- ▶ A way to quantify I/O cost on a DAG (why a DAG?)
- ▶ “Red Hot” pebbles: data that can be computed on
- ▶ “Blue Cool” pebbles: data that is stored, but not available for computation without I/O

Game Setup:

- ▶ Place blue pebbles on inputs

One ‘turn’:

- ▶ Swap red \rightarrow blue or blue \rightarrow red
- ▶ Place red pebble on successor if all predecessors have red
- ▶ Remove any pebble

Note: Can allow “Red/Purple/Blue/Black”: more levels

[[Hong/Kung '81](#)]

Hong/Kung: Red/Blue Pebble Game

Q: What are the cost metrics in this model?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the question above.

What does a Hong/Kung lower bound for Matrix-Matrix look like?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the question above.

Cache-Oblivious Algorithms

Annoying chore: Have to pick multiple machine-adapted block sizes in cache-adapted algorithms, one for each level in the memory hierarchy, starting with registers.

Idea:

- ▶ Step 1: Express algorithm recursively in divide & conquer manner
- ▶ Step 2: Pick a strategy to decrease block size
Give examples of block size strategies, e.g. for MMM:



Result:

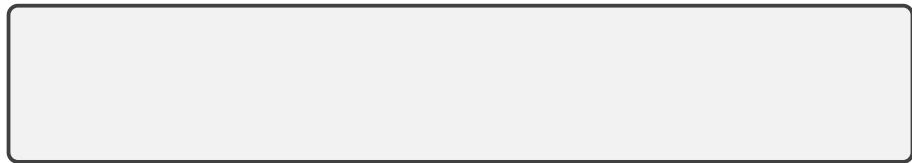
- ▶ Asymptotically optimal on Hong/Kung metric

Cache-Oblivious Algorithms: Issues

Observed results?

A large, empty rectangular box with a thin black border and rounded corners, intended for notes or results.

What are potential issues on actual hardware?

A large, empty rectangular box with a thin black border and rounded corners, intended for notes or results.

[[Yotov et al. '07](#)]

Recall: Big-O Notation

Classical Analysis of Algorithms (e.g.):

$$\text{Cost}(n) = O(n^3).$$

Precise meaning? Anything missing from that statement?



Comment: “Asymptotically Optimal”

Comments on asymptotic statements about cost in relation to high performance?

- ▶ No statement about finite n
- ▶ No statement about the constant

Net effect: Having an understanding of asymptotic cost is *necessary*, but *not sufficient* for high performance.

HPC is in the business of minimizing C in:

$$\text{Cost}(n) \leq C \cdot n^3 \quad (\text{for all } n)$$

Alignment

Alignment describes the process of matching the base address of:

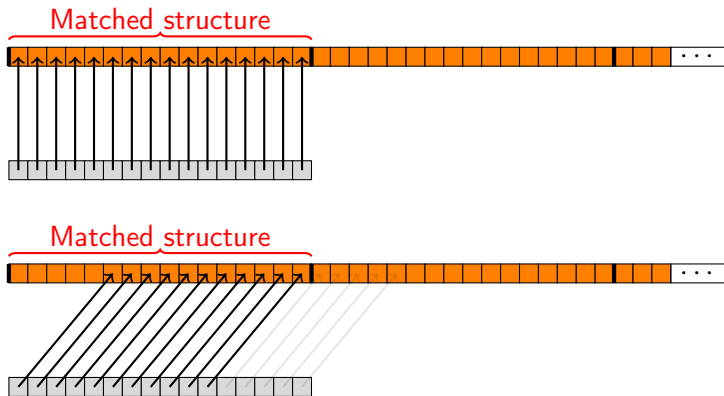
- ▶ Single word: double, float
- ▶ SIMD vector
- ▶ Larger structure

To machine granularities:



Q: What is the performance impact of misalignment?

Performance Impact of Misalignment



SIMD: Basic Idea

What's the basic idea behind SIMD?



What architectural need does it satisfy?



Typically characterized by width of data path:

- ▶ SSE: 128 bit (4 floats, 2 doubles)
- ▶ AVX-2: 256 bit (8 floats, 4 doubles)
- ▶ AVX-512: 512 bit (16 floats, 8 doubles)

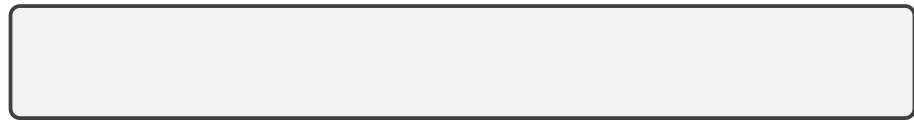
SIMD: Architectural Issues

Realization of inter-lane comm. in SIMD? Find instructions. ([Intel Intrinsic Guide](#))

Name tricky/inefficient aspects in terms of expressing SIMD:

SIMD: Transposes

Why are transposes important? Where do they occur?



Example implementation aspects:

- ▶ HPTT: [[Springer et al. '17](#)]
- ▶ github: [springer13/hptt](#) [8x8 transpose microkernel](#)
- ▶ Q: Why 8x8?

Outline

Introduction

Notes

Notes (unfilled, with empty boxes)

Notes (source code on Github)

About This Class

Why Bother with Parallel Computers?

Lowest Accessible Abstraction: Assembly

Architecture of an Execution Pipeline

Architecture of a Memory System

Shared-Memory Multiprocessors

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance Oriented Languages and Abstractions

Multiple Cores vs Bandwidth

Assume:

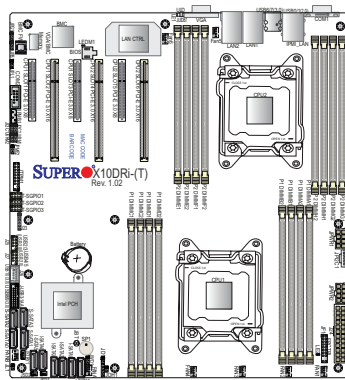
- ▶ memory latency of 100 ns
- ▶ peak DRAM bandwidth of 100 GB/s (per socket)

How many cache lines should be/are in flight at one time?



[[McCalpin '18](#)]

Topology and NUMA



[SuperMicro Inc. '15]

Demo:

- ▶ Show 1stopo on porter, from [hwloc](#).
- ▶ [1stopo on MI300](#)

Placement and Pinning

Who decides on what core my code runs? How?



Who decides on what NUMA node memory is allocated?



[Demo: intro/NUMA and Bandwidths](#)

What is the main expense in NUMA?



Cache Coherence

What is *cache coherence*?



How is cache coherence implemented?



[Cache coherence simulation](#)

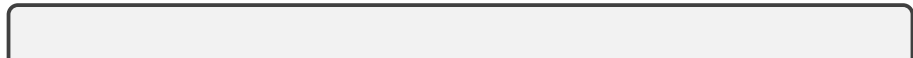
What are the performance impacts?

- ▶ [Demo: intro/Threads vs Cache](#)
- ▶ [Demo: intro/Lock Contention](#)

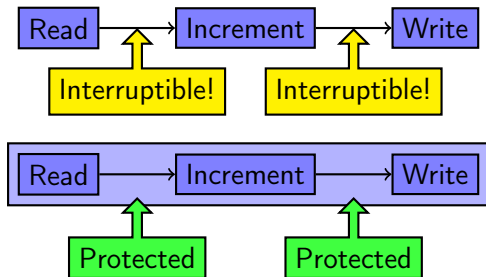
MESI Sequence to Try



Simplifications in the model?



'Conventional' vs Atomic Memory Update



Outline

Introduction

Machine Abstractions

C

OpenCL/CUDA

Convergence, Differences in Machine Mapping

Lower-Level Abstractions: SPIR-V, PTX

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Outline

Introduction

Machine Abstractions

C

OpenCL/CUDA

Convergence, Differences in Machine Mapping

Lower-Level Abstractions: SPIR-V, PTX

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Atomic Operations: Compare-and-Swap

```
#include <stdatomic.h>
_Bool atomic_compare_exchange_strong(
    volatile A* obj, C* expected, C desired );
```

What does volatile mean?

What does this do?

How might you use this to implement atomic FP multiplication?

Memory Ordering

Why is Memory Ordering a Problem?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

What's the purpose of different memory orders?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

Memory ordering semantics



Example: A Semaphore With Atomics

```
#include <stdatomic.h> // mo_→memory_order, a_→atomic
typedef struct { atomic_int v;} naive_sem_t;
void sem_down(naive_sem_t *s)
{
    while (1) {
        while (a_load_explicit(&(s->v), mo_____) < 1)
            spinloop_body();
        int tmp=a_fetch_add_explicit(&(s->v), -1, mo_____rel);
        if (tmp >= 1)
            break; // we got the lock
        else // undo our attempt
            a_fetch_add_explicit(&(s->v), 1, mo_____);
    }
}
void sem_up(naive_s_t *s) {
    a_fetch_add_explicit(&(s->v), 1, mo_____);
}
```

C: What is 'order'?

C11 Committee Draft, December '10, Sec. 5.1.2.3, "Program execution":

- ▶ (3) *Sequenced before* is an asymmetric, transitive, pair-wise relation between evaluations executed by a single thread, which induces a partial order among those evaluations. Given any two evaluations A and B, if A is sequenced before B, then the execution of A shall precede the execution of B. (Conversely, if A is sequenced before B, then B is sequenced after A.) If A is not sequenced before or after B, then A and B are unsequenced. Evaluations A and B are *indeterminately sequenced* when A is sequenced either before or after B, but it is unspecified which. The presence of a sequence point between the evaluation of expressions A and B implies that every value computation and side effect associated with A is sequenced before every value computation and side effect associated with B. (A summary of the *sequence points* is given in annex C.)

Q: Where is this definition used (in the standard document)?

C: What is 'order'? (Encore)

C11 Draft, 5.1.2.4 "Multi-threaded executions and data races":

- ▶ All modifications to a particular atomic object M occur in some particular total order, called the *modification order* of M.
- ▶ An evaluation A *carries a dependency* to an evaluation B if ...
- ▶ An evaluation A is *dependency-ordered* before an evaluation B if...
- ▶ An evaluation A *inter-thread happens before* an evaluation B if...
- ▶ An evaluation A *happens before* an evaluation B if...

Why is this so subtle?



C: How Much Lying is OK?

C11 Committee Draft, December '10, Sec. 5.1.2.3, "Program execution":

- ▶ (1) The semantic descriptions in this International Standard describe the behavior of an abstract machine in which issues of optimization are irrelevant.
- ▶ (2) Accessing a volatile object, modifying an object, modifying a file, or calling a function that does any of those operations are all *side effects*, which are changes in the state of the execution environment.
[...]

C: How Much Lying is OK?

- ▶ (4) In the abstract machine, all expressions are evaluated as specified by the semantics. An actual implementation need not evaluate part of an expression if it can deduce that its value is not used and that no needed side effects are produced (including any caused by calling a function or accessing a volatile object).
- ▶ (6) The least requirements on a conforming implementation are:
 - ▶ Accesses to volatile objects are evaluated strictly according to the rules of the abstract machine.
 - ▶ At program termination, all data written into files shall be identical to the result that execution of the program according to the abstract semantics would have produced.
 - ▶ The input and output dynamics of interactive devices shall take place as specified in 7.21.3. The intent of these requirements is that unbuffered or line-buffered output appear as soon as possible, to ensure that prompting messages actually appear prior to a program waiting for input.

This is the observable behavior of the program.

Arrays

Why are **arrays** the dominant data structure in high-performance code?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

Any comments on C's arrays?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

Arrays vs Abstraction

Arrays-of-Structures or Structures-of-Arrays? Give an example.

Language aspects of the distinction? Salient example?

C and Multi-Dimensional Arrays: A Saving Grace

// YES:

```
void f(int m, int n, double (*)[m][n]);
```

// NO:

```
struct ary {  
    int m;  
    int n;  
    double (*array)[m][n];  
};
```

// YES:

```
struct ary {  
    int m;  
    int n;  
    double a[];  
};
```

SIMD

Name language mechanisms for SIMD:



[Demo: machabstr/Ways to SIMD](#)

Outer-Loop/inner-Loop Vectorization

Contrast *outer-loop* vs *inner-loop* vectorization.



Side q: Would you consider GPUs outer- or inner-loop-vectorizing?

Alignment: How?

The old way:

```
int __attribute__((aligned (8))) a_int;
```

Difference between these two?

```
int __attribute__((aligned (8))) * ptr_t_1;  
int * __attribute__((aligned (8))) ptr_t_2;
```

The 'new' way (C/C++11):

```
struct alignas(64) somestruct_t { /* ... */ };  
struct alignas(alignof(other_t))  
    somestruct_t { /* ... */ };  
struct  
    alignas(  
        std::hardware_destructive_interference_size)  
    somestruct_t { /* ... */ };
```

What is *constructive interference*?

Alignment: Why?

What is the concrete impact of the constructs on the previous slide?



Pointers and Aliasing

[Demo: machabstr/Pointer Aliasing](#)

Register Pressure

What if the register working set gets larger than the registers can hold?
What is the performance impact?



[Demo: machabstr/Register Pressure](#)

Object-Oriented Programming

Object-oriented programming: The weapon of choice for encapsulation and separation of concerns!

Performance perspective on OOP?



[Demo: machabstr/Object Orientation vs Performance](#)

Being Nice to Your Compiler

Some rules of thumb:

- ▶ Use indices rather than pointers
- ▶ Extract common subexpressions
- ▶ Make functions static
- ▶ Use `const`
- ▶ Avoid store-to-load dependencies

What are the concrete impacts of doing these things?

Outline

Introduction

Machine Abstractions

C

OpenCL/CUDA

Convergence, Differences in Machine Mapping

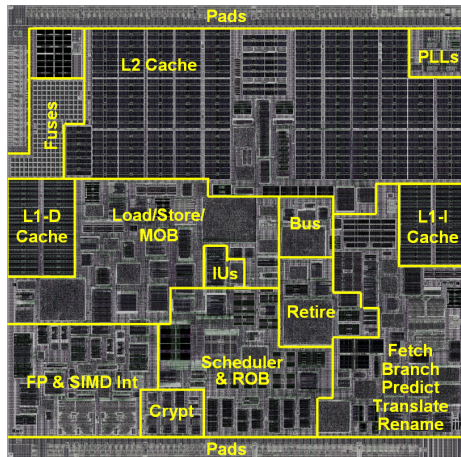
Lower-Level Abstractions: SPIR-V, PTX

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

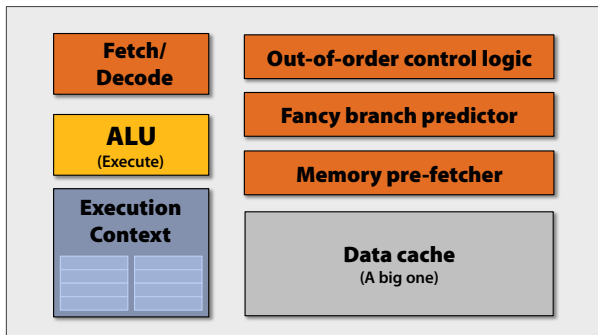
Polyhedral Representation and Transformation

Chip Real Estate



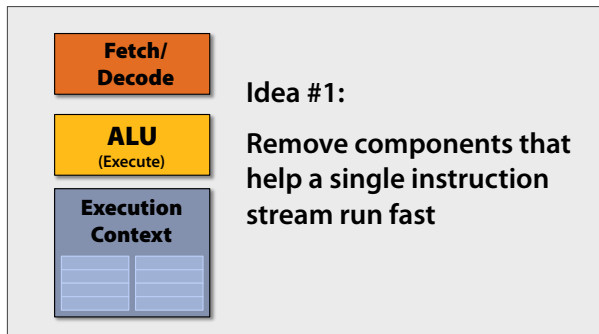
Die floorplan: VIA Isaiah (2008).
65 nm, 4 SP ops at a time, 1 MiB L2.

“CPU-style” Cores



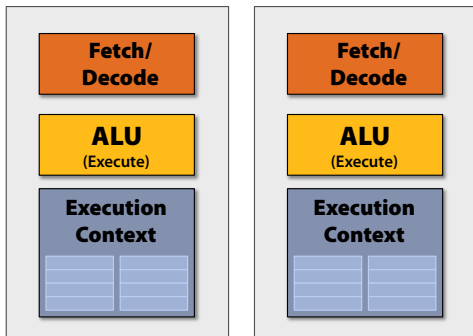
[Fatahalian '08]

Slimming down



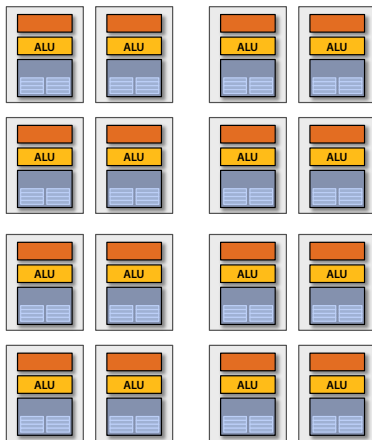
[Fatahalian '08]

More Space: Double the Number of Cores



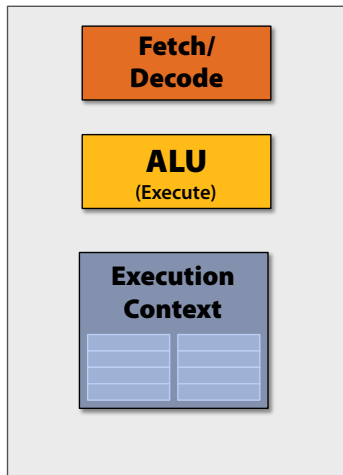
[Fatahalian '08]

Even more



[Fatahalian '08]

SIMD

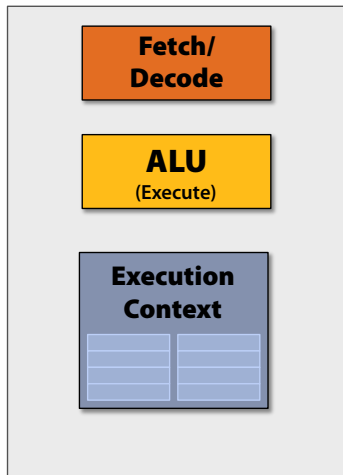


Idea #2: SIMD

Amortize cost/complexity of managing an instruction stream across many ALUs

[Fatahalian '08]

SIMD

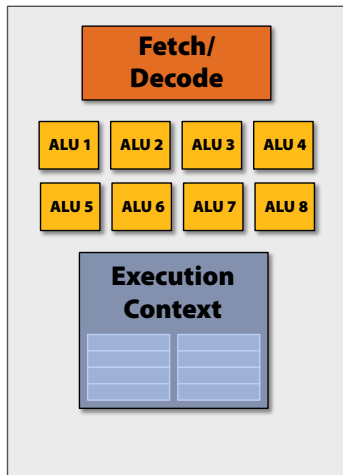


Idea #2: SIMD

Amortize cost/complexity of managing an instruction stream across many ALUs

[Fatahalian '08]

SIMD

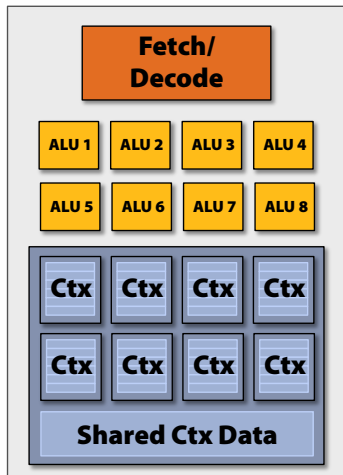


Idea #2: SIMD

Amortize cost/complexity of managing an instruction stream across many ALUs

[Fatahalian '08]

SIMD



Idea #2: SIMD

Amortize cost/complexity of managing an instruction stream across many ALUs

[Fatahalian '08]

Latency Hiding

- ▶ Latency (mem, pipe) hurts non-OOO cores
- ▶ Do *something* while waiting

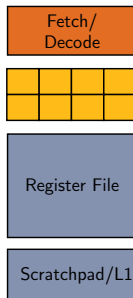
What is the unit in which work gets scheduled on a GPU?



How can we keep busy?



Before:



After:

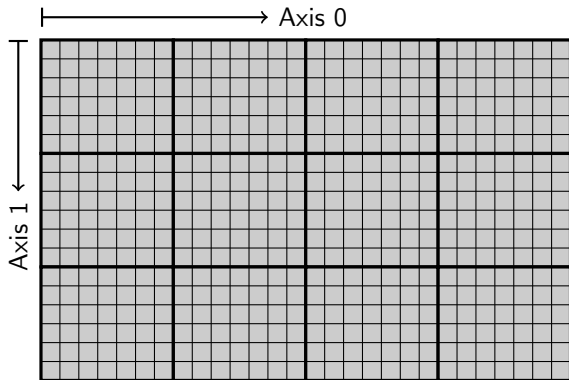


GPUs: Core Architecture Ideas

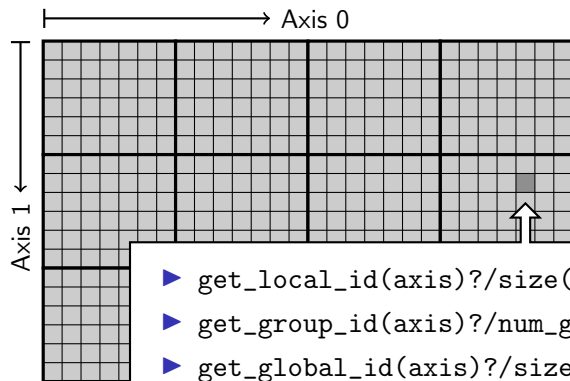
Three core ideas:



'SIMT'



Wrangling the Grid

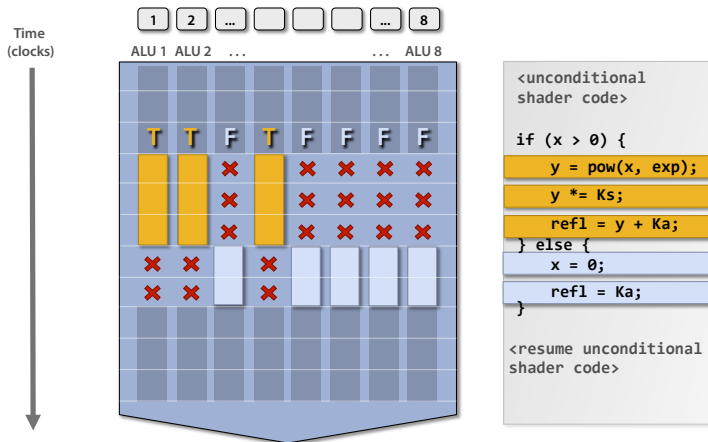


- ▶ `get_local_id(axis)?/size(axis)?`
 - ▶ `get_group_id(axis)?/num_groups(axis)?`
 - ▶ `get_global_id(axis)?/size(axis)?`
- `axis=0,1,2,...`

Demo CL code

[Demo: machabstr/Hello GPU](#)

'SIMT' and Branches



[Fatahalian '08]

GPU Abstraction: Core Model Ideas

How do these aspects show up in the model?



GPU Address Spaces

Hardware	CL adjective	OpenCL	CUDA
SIMD lane	private	Work Item	Thread
SIMD Vector	—	Subgroup	Warp
Core	local	Workgroup	Thread Block
Processor	global	NDRange	Grid

GPU: Communication

What forms of communication exist at each scope?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the question above.

Can we just do locking like we might do on a CPU?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide an answer to the question above.

GPU Programming Model: Commentary

Advantages:

A large, empty rectangular box with a thin black border, intended for listing the advantages of the GPU programming model.

Disadvantages:

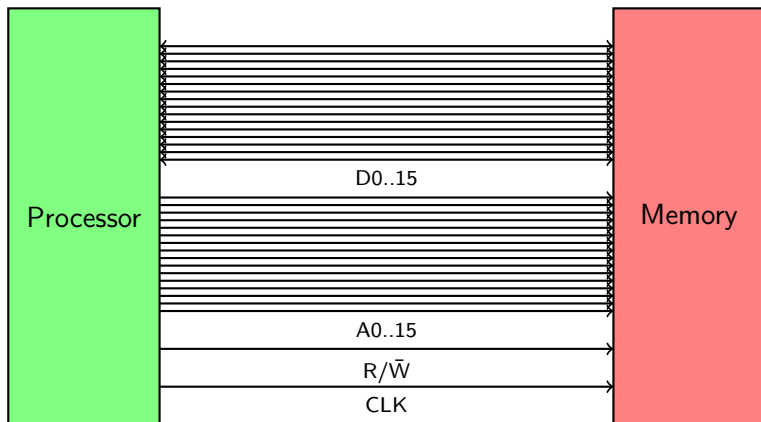
A large, empty rectangular box with a thin black border, intended for listing the disadvantages of the GPU programming model.

Performance: Limits to Concurrency

What limits the amount of concurrency exposed to GPU hardware?



Memory Systems: Recap



Parallel Memories

Problem: Memory chips have only one data bus.

So how can multiple threads read multiple data items from memory simultaneously?



Where does banking show up?



Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank ↑	2	6	10	14	18	22	...
	1	5	9	13	17	21	...
	0	4	8	12	16	20	...
	→ Address						

3

2

Thread
↑

1

0

Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank

2	6	10	14	18	22	...
---	---	----	----	----	----	-----

1	5	9	13	17	21	...
---	---	---	----	----	----	-----

0	4	8	12	16	20	...
---	---	---	----	----	----	-----

→ Address

`local_variable[lid(0)]`

3

2

Thread

1

0

Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank	2	6	10	14	18	22	...
------	---	---	----	----	----	----	-----

1	5	9	13	17	21	...
---	---	---	----	----	----	-----

0	4	8	12	16	20	...
---	---	---	----	----	----	-----

→ Address

3

2

Thread

1

0

```
local_variable[BANK_COUNT*lid(0)]
```

Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank ↑	2	6	10	14	18	22	...
	1	5	9	13	17	21	...
	0	4	8	12	16	20	...
	→						Address

3

2

Thread

1

0

```
local_variable[(BANK_COUNT+1)*lid(0)]
```

Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank	2	6	10	14	18	22	...
	1	5	9	13	17	21	...
	0	4	8	12	16	20	...
	→ Address						

3

2

Thread

1

0

```
local_variable[ODD_NUMBER*lid(0)]
```

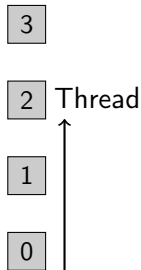
Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank	2	6	10	14	18	22	...
	1	5	9	13	17	21	...
	0	4	8	12	16	20	...
	→ Address						

`local_variable[2*lid(0)]`



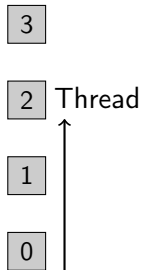
Memory Banking

Fill in the access pattern:

3	7	11	15	19	23	...
---	---	----	----	----	----	-----

Bank ↑	2	6	10	14	18	22	...
	1	5	9	13	17	21	...
	0	4	8	12	16	20	...
	→						Address

```
local_variable[f(gid(0))]
```



Memory Banking: Observations

- ▶ Factors of two in the stride: generally bad
- ▶ In a conflict-heavy access pattern, padding can help
 - ▶ Usually not a problem since scratchpad is transient by definition
- ▶ Word size (bank offset) may be adjustable (Nvidia)

Given that unit strides are beneficial on global memory access, how do you realize a transpose?



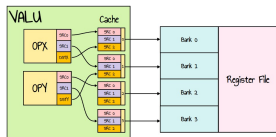
Memory Banking: AMD RDNA3 Registers

Manual dual-issue:

OpCodeX DSTX, SRCX0, SRCX1

:: OpCodeY DSTY, SRCY0, SRCY1

- ▶ Insns must be independent
- ▶ SRCX0 and SRCY0 must use different VGPR banks
- ▶ Dest VGPRs: one must be even and the other odd

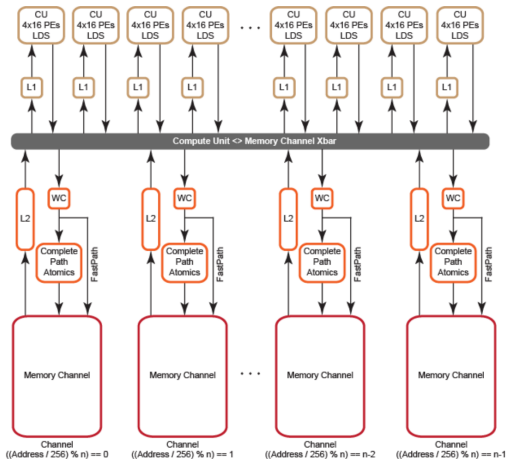


RDNA3 specific:

- ▶ There are 4 VGPR banks (indexed by SRC[1:0]), and each bank has a cache.
- ▶ Each cache has 3 read ports: one dedicated to SRC0, one dedicated to SRC1 and one for SRC2.
- ▶ A cache can read all 3 of them at once, but it can't read two SRC0's at once (or SRC1/2).

[[Vince '25](#)]

GPU Global Memory System



[GCN Optimization Manual, AMD](#)

GPU Global Memory Channel Map: Example

Byte address decomposition:

Address											Bank	Chnl	Address							
31											?	11		108		7				0

Implications:

- ▶ Transfers between compute unit and channel have granularity
 - ▶ Reasonable guess: warp/wavefront size \times 32bits
 - ▶ Should strive for good utilization ('*Coalescing*')
- ▶ Channel count often *not* a power of two \rightarrow complex mapping
 - ▶ *Channel conflicts* possible
- ▶ Also *banked*
 - ▶ *Bank conflicts* also possible

GPU Global Memory: Performance Observations

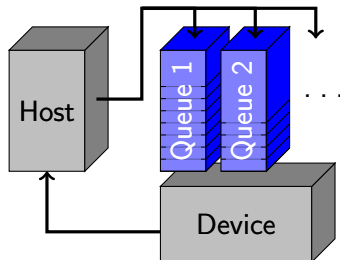
Key quantities to observe for GPU global memory access:

Are there any guaranteed-good memory access patterns?

- ▶ Need to consider access pattern *across entire device*
- ▶ *GPU caches*: Use for *spatial*, not for temporal locality
- ▶ Switch available: L1/Scratchpad partitioning
 - ▶ Settable on a per-kernel basis
- ▶ Since GPUs have meaningful caches at this point:
Be aware of cache annotations (see later)

Host-Device Concurrency

- ▶ Host and Device run asynchronously
- ▶ Host submits to queue:
 - ▶ Computations
 - ▶ Memory Transfers
 - ▶ Sync primitives
 - ▶ ...
 - ▶ *Batches* of these
 - ▶ *Mutable* batches of these
 - ▶ Nvidia: "CUDA Graphs"
 - ▶ OpenCL: "Command buffers"
- ▶ Host can wait for:
 - ▶ *drained* queue
 - ▶ Individual "events"
- ▶ Profiling



Timing GPU Work

How do you find the execution time of a GPU kernel?

A large, empty rectangular box with a thin black border, intended for the user to provide an answer to the question above.

How do you do this asynchronously?

A large, empty rectangular box with a thin black border, intended for the user to provide an answer to the question above.

Host-Device Data Exchange

Sad fact: Must get data onto device to compute

- ▶ Transfers can be a bottleneck
- ▶ If possible, overlap with computation
- ▶ Pageable memory incurs difficulty in GPU-host transfers, often entails (another!) CPU side copy
- ▶ “Pinned memory”: unpageable, avoids copy
 - ▶ Various *system-defined* ways of allocating pinned memory

“Unified memory” (CUDA)/“Shared Virtual Memory” (OpenCL):

- ▶ GPU directly accesses host memory
- ▶ Main distinction: Coherence
 - ▶ “Coarse grain”: Per-buffer fences
 - ▶ “Fine grain buffer”: Byte-for-byte coherent (device mem)
 - ▶ “Fine grain system”: Byte-for-byte coherent (anywhere)

Performance: Ballpark Numbers?

Bandwidth host/device:

Bandwidth on device (A100 PCIe):

Flop throughput? (A100 PCIe)

Kernel launch overhead?

Good source of details: [Wikipedia: List of Nvidia GPUs](#)

Outline

Introduction

Machine Abstractions

C

OpenCL/CUDA

Convergence, Differences in Machine Mapping

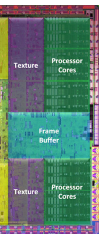
Lower-Level Abstractions: SPIR-V, PTX

Performance: Expectation, Experiment, Observation

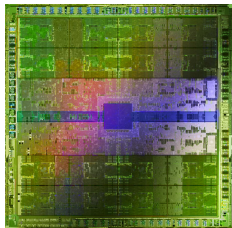
Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

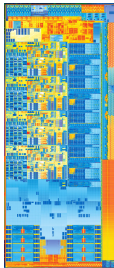
Die Shot Gallery (old)



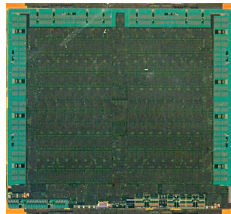
200
(8)



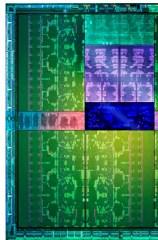
Nv Fermi
(2010)



Intel IVB
(2012)

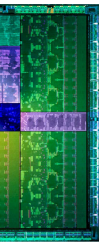


AMD Tahiti
(2012)

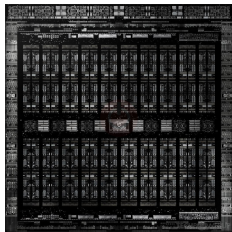


Nv GK110
(2012)

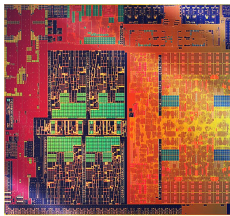
Die Shot Gallery (new)



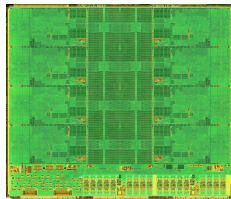
AMD Zen 4
(2022)



Nv GV100
(2018)



Intel ICL
(2018)



AMD Zen 5
(2024)



AMD MI300
(2024)

Trends in Processor Architecture

What can we expect from future processor architectures?



Common Challenges

What are the common challenges encountered by both CPUs and GPUs?



Goal: Try to see CPUs and GPUs as points in a design space 'continuum' rather than entirely different things.

Outline

Introduction

Machine Abstractions

C

OpenCL/CUDA

Convergence, Differences in Machine Mapping

Lower-Level Abstractions: SPIR-V, PTX

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

PTX: Demo

Demo: machabstr/PTX and SASS

Nvidia PTX manual

PTX: Cache Annotations

Loads:

- .ca Cache at all levels—likely to be accessed again
- .cg Cache at global level (cache in L2 and below and not L1)
- .cs Cache streaming—likely to be accessed once
- .lu Last use
- .cv Consider cached system memory lines stale—fetch again

Stores:

- .wb Cache write-back all coherent levels
- .cg Cache at global level (cache in L2 and below and not L1)
- .cs Cache streaming—likely to be accessed once
- .wt Cache write-through (to system memory)

Lost/hidden at the C level!

SPIR-V

Currently: C (OpenCL C, GLSL, HLSL) used as intermediate representations to feed GPUs.

Downsides:

- ▶ Compiler heuristics may be focused on human-written code
- ▶ Parsing overhead (preprocessor!)
- ▶ C semantics may not match (too high-level)

SPIR-V:

- ▶ Goal: Common intermediate representation (“IR”) for all GPU-facing code (Vulkan, OpenCL)
- ▶ “Extended Instruction Sets”:
 - ▶ General compute (OpenCL/CUDA) needs: pointers, special functions
- ▶ Different from “SPIR” (tweaked LLVM IR)

SPIR-V Example

```
%2 = OpTypeVoid
%3 = OpTypeFunction %2                ; void ()
%6 = OpTypeFloat 32                  ; 32-bit float
%7 = OpTypeVector %6 4                ; vec4
%8 = OpTypePointer Function %7       ; function-local vec4*
%10 = OpConstant %6 1
%11 = OpConstant %6 2
%12 = OpConstantComposite %7 %10 %10 %11 %10 ; vec4(1.0, 1.0, 2.0, 1.0)
%13 = OpTypeInt 32 0                  ; 32-bit int, sign-less
%14 = OpConstant %13 5
%15 = OpTypeArray %7 %14
```

[...]

```
%34 = OpLoad %7 %33
%38 = OpAccessChain %37 %20 %35 %21 %36 ; s.v[2]
%39 = OpLoad %7 %38
%40 = OpFAdd %7 %34 %39
      OpStore %31 %40
      OpBranch %29
%41 = OpLabel                        ; else
%43 = OpLoad %7 %42
%44 = OpExtInst %7 %1 Sqrt %43       ; extended instruction sqrt
%45 = OpLoad %7 %9
%46 = OpFMul %7 %44 %45
      OpStore %31 %46
```

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

- Forming Expectations of Performance

- Timing Experiments and Potential Issues

- Profiling and Observable Quantities

- Practical Tools: perf, toplev, likwid

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Forming Expectations of Performance

Timing Experiments and Potential Issues

Profiling and Observable Quantities

Practical Tools: perf, toplev, likwid

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Qualifying Performance

- ▶ What is *good* performance?
- ▶ Is speed-up (e.g. GPU vs CPU? C vs Matlab?) a meaningful way to assess performance?
- ▶ How else could one *form an understanding* of performance?

[Hager et al. '17](#)

[Hockney et al. '89](#)

A Story of Bottlenecks

Imagine:

- ▶ A bank with a few service desks
- ▶ A revolving door at the entrance

What situations can arise at *steady-state*?

A large, empty rectangular box with a thin black border, intended for the user to write their answer to the question about steady-state situations.

What numbers do we need to characterize performance of this system?

A large, empty rectangular box with a thin black border, intended for the user to write their answer to the question about performance numbers.

A Story of Bottlenecks (cont'd)

- ▶ P_{peak} : [task/sec] Peak throughput of the service desks
- ▶ I : [tasks/customer] Intensity
- ▶ b : [customers/sec] Throughput of the revolving door

What is the aggregate throughput?



Hager et al. '17

Application in Computation

Translate the bank analogy to computers:

Which parts of this are task-dependent?

$$P \leq \min(P_{\text{peak}}, I \cdot b)$$

A Graphical Representation: 'Roofline'

Plot (often log-log, but not necessarily):

- ▶ X-Axis: Intensity
- ▶ Y-Axis: Performance

What does our inequality correspond to graphically?

$$P \leq \min(P_{\text{peak}}, I \cdot b)$$



What does the shaded area mean?



Example: Vector Addition

```
double r, s, a[N];  
for (i=0; i<N; ++i)  
    a[i] = r + s * a[i];}
```

Find the parameters and make a prediction.

performance ↑

Refining the Model

- ▶ P_{\max} : Applicable peak performance of a loop, assuming that data comes from the fastest data path (this is not necessarily P_{peak})
- ▶ Computational intensity (“work” per byte transferred) over the slowest data path utilized
- ▶ b : Applicable peak bandwidth of the slowest data path utilized


[Hager et al. '17](#)

Calibrating the Model: Bandwidth

Typically done with the STREAM benchmark.

Four parts: Copy, Scale, Add, Triad $a[i] = b[i] + s \cdot c[i]$

Do the four measurements matter?



Any pitfalls?



[McCalpin: STREAM](#)

Calibrating the Model: Peak Throughput

Name aspects that should/could be factored in when determining peak performance:



Practical Tool: IACA

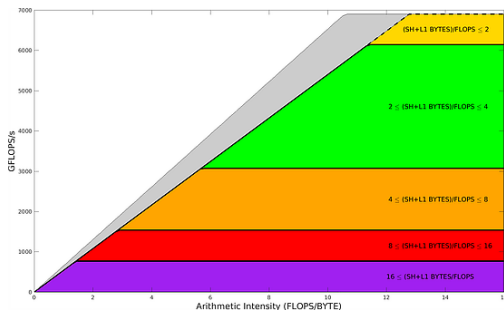
Question: Where to obtain an estimate of P_{\max} ?

Demo: [perf/Forming Architectural Performance Expectations](#)

Questions:

- ▶ What does IACA do about memory access? / the memory hierarchy?

An Example: Exploring Titan V Limits



- ▶ Memory bandwidth: 652 GB/s theoretical, 540 GB/s achievable
- ▶ Scratchpad / L1 throughput:
80 (cores) × 32 (simd width) × 4 (word bytes) × 1.2 (base clock) ≈ 12.288 TB/s
- ▶ Theoretical peak flops of 6.9 TFLOPS/s [Wikipedia]

Warburton '18

Rooflines: Assumptions

What assumptions are built into the roofline model?



Important to remember:

- ▶ It is what you make of it—the better your calibration, the more info you get
- ▶ But: Calibrating on experimental data loses predictive power (e.g. SPMV)

Modeling Parallel Speedup: A 'Universal' Scaling Law

Develop a model of *throughput* $X(N)$ for a given load N , assuming execution resources scale with N .



[[Gunther '93](#)]

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Forming Expectations of Performance

Timing Experiments and Potential Issues

Profiling and Observable Quantities

Practical Tools: perf, toplev, likwid

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Combining Multiple Measurements

How can one combine multiple performance measurements? (e.g. “average speedup”?)

Example: Which computer should you buy?

Execution time [s]	Computer A	Computer B	Computer C
Program 1	1	10	20
Program 2	1000	100	20

Combining Multiple Measurements: Observations

	Computer A	Computer B	Computer C
Arithmetic mean	500.5	55	20
Geometric mean	31.622	31.622	20

Timing Experiments: Pitfalls

What are potential issues in timing experiments? (What can you do about them?)

A large, empty rectangular box with a black border, intended for a response.

Timing Experiments: Pitfalls (part 2)

What are potential issues in timing experiments? (What can you do about them?)



Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Forming Expectations of Performance

Timing Experiments and Potential Issues

Profiling and Observable Quantities

Practical Tools: perf, toplev, likwid

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Profiling: Basic Approaches

Measurement of “quantities” relating to performance

- ▶ Exact: Through binary instrumentation (valgrind/Intel Pin/...)
- ▶ Sampling: At *some* interval, examine the program state

We will focus on profiling by *sampling*.

Big questions:

- ▶ What to measure?
- ▶ At what intervals?

Defining Intervals: Performance Counters

A *performance counter* is a counter that increments every time a given **event** occurs.

What events?

- ▶ [Demo: perf/Using Performance Counters](#)
- ▶ see also [Intel SDM, Volume 3](#)

Interaction with performance counters:

- ▶ Read repeatedly from user code
- ▶ Interrupt program execution when a threshold is reached
- ▶ Limited resource!
 - ▶ Only a few available: 4-8 per core
 - ▶ Each can be configured to count one type of event
 - ▶ **Idea:** Alternate counter programming at some rate (requires steady-state execution!)

Profiling: What to Measure

- ▶ Raw counts are hard to interpret
- ▶ Often much more helpful to look at *ratios* of counts per core/subroutine/loop/...

What ratios should one look at?

[Demo: perf/Using Performance Counters](#)

Profiling: Useful Ratios

Basic examples:

- ▶ $(\text{Events in Routine 1})/(\text{Events in Routine 2})$
- ▶ $(\text{Events in Line 1})/(\text{Events in Line 2})$
- ▶ $(\text{Count of Event 1 in X})/(\text{Count of Event 2 in X})$

Architectural examples:



Issue with 'instructions' as a metric?



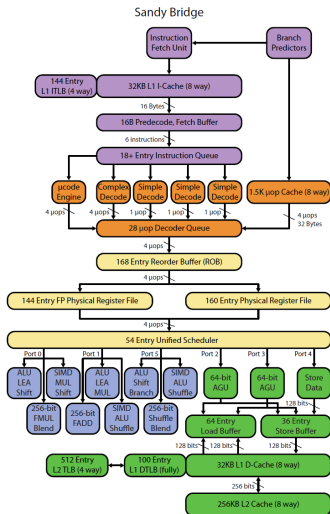
“Top-Down” Performance Analysis

Idea: Account for useful work per available issue slot
What is an issue slot?



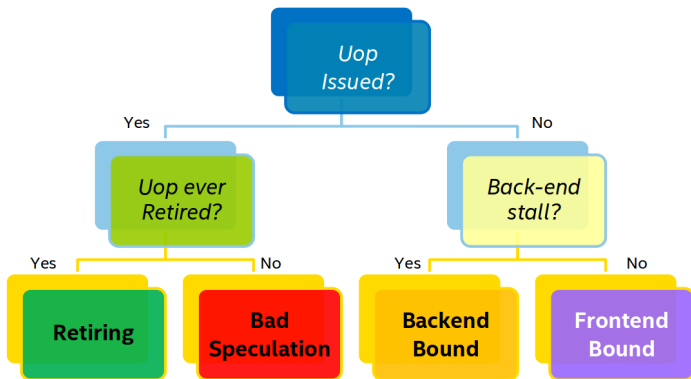
[[Yasin '14](#)]

Issue Slots: Recap



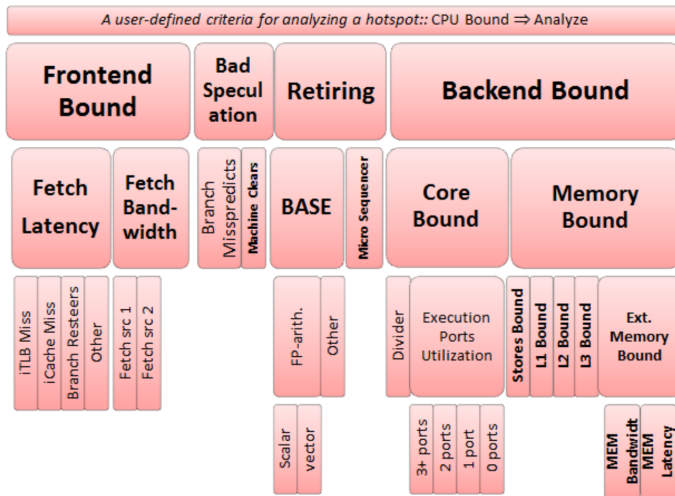
[David Kanter / Realworldtech.com]

What can happen to an issue slot: at a high level?



[Yasin '14]

What can happen to an issue slot: in detail?



[Yasin '14]

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Forming Expectations of Performance

Timing Experiments and Potential Issues

Profiling and Observable Quantities

Practical Tools: perf, toplev, likwid

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Demo: Performance Counters

Show the rest of:

[Demo: perf/Using Performance Counters](#)

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

- Expression Trees

- Parallel Patterns and Array Languages

Polyhedral Representation and Transformation

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Expression Trees

Parallel Patterns and Array Languages


Polyhedral Representation and Transformation

Expression Trees and Term Rewriting

Demos:

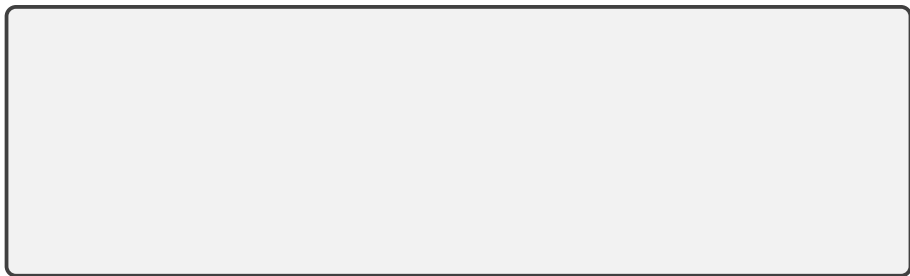
- ▶ [Demo: lang/01 Expression Trees](#)
- ▶ [Demo: lang/02 Traversing Trees](#)
- ▶ [Demo: lang/03 Defining Custom Node Types](#)
- ▶ [Demo: lang/04 Common Operations](#)

How do expression trees come to be? (not our problem here)



Embedded languages

Main challenge: Obtaining a syntax tree. Approaches?



Macros: Goals and Approaches

What is a macro?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question "What is a macro?".

What data do macro systems operate on?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question "What data do macro systems operate on?".

Macros: Textual and Syntactic, Hygiene

Macros: What can go wrong if you're not careful?

```
#define INCI(i) do { int a=0; ++i; } while(0)
int main(void)
{
    int a = 4, b = 8;
    INCI(a);
    INCI(b);
    printf("a is now %d, b is now %d\n", a, b);
    return 0;
}
```

How can the problem above be avoided?

Towards Execution

[Demo: lang/06 Towards Execution](#)

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Expression Trees

Parallel Patterns and Array Languages

Polyhedral Representation and Transformation

Reduction

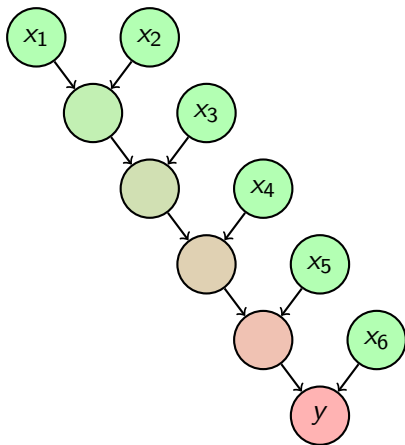
$$y = f(\cdots f(f(x_1, x_2), x_3), \cdots, x_N)$$

where N is the input size.

Also known as

- ▶ Lisp/Python function `reduce` (Scheme: `fold`)
- ▶ C++ STL `std::accumulate`

Reduction: Graph



Approach to Reduction



Can we do better?

“Tree” very imbalanced. What property of f would allow ‘rebalancing’?

$$f(f(x, y), z) = f(x, f(y, z))$$

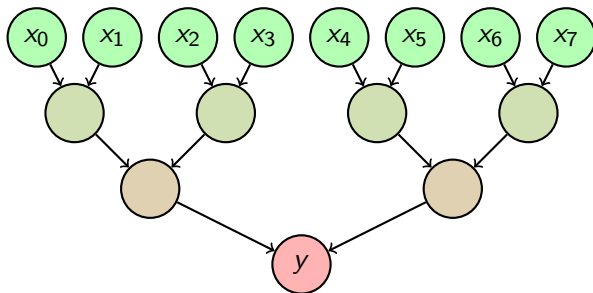
Looks less improbable if we let

$x \circ y = f(x, y)$:

$$x \circ (y \circ z) = (x \circ y) \circ z$$

Has a very familiar name: *Associativity*

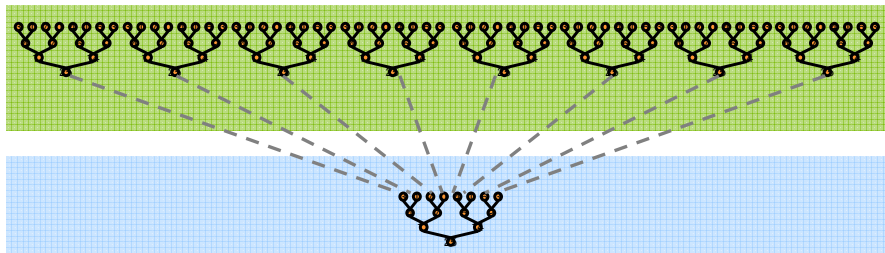
Reduction: A Better Graph



Processor allocation?

Mapping Reduction to SIMD/GPU

- ▶ Obvious: Want to use tree-based approach.
- ▶ Problem: Two scales, Work group and Grid
 - ▶ to occupy both to make good use of the machine.
- ▶ In particular, need synchronization after each tree stage.
- ▶ Solution: Use a two-scale algorithm.



In particular: Use multiple grid invocations to achieve inter-workgroup synchronization.

But no. Not even close.

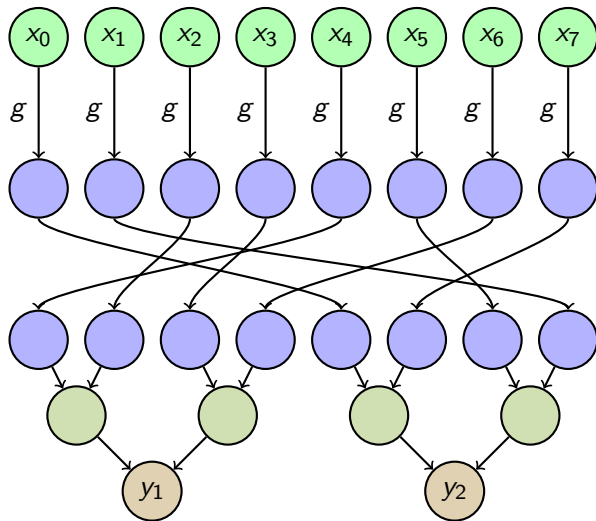
Sounds like this:

$$y = f(\cdots f(f(g(x_1), g(x_2)), \\ g(x_3)), \cdots, g(x_N))$$

where N is the input size.

- ▶ Lisp naming, again
- ▶ Mild generalization of reduction

Map-Reduce: Graph



Scan

$$y_1 = x_1$$

$$y_2 = f(y_1, x_2)$$

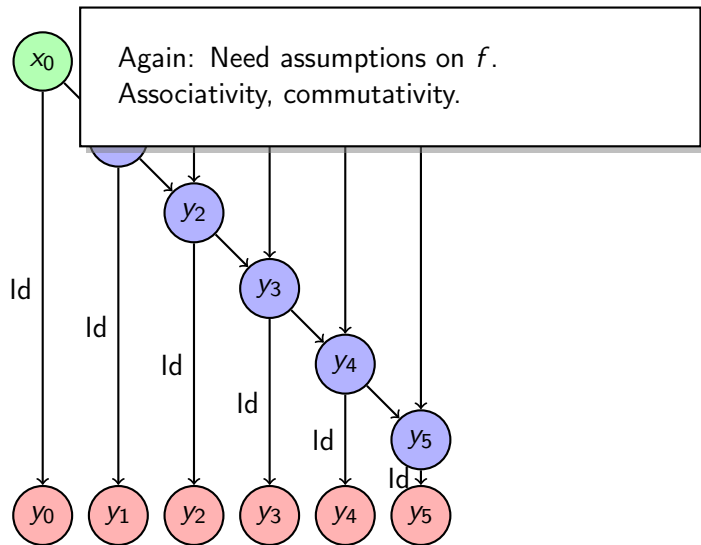
$$\vdots = \vdots$$

$$y_N = f(y_{N-1}, x_N)$$

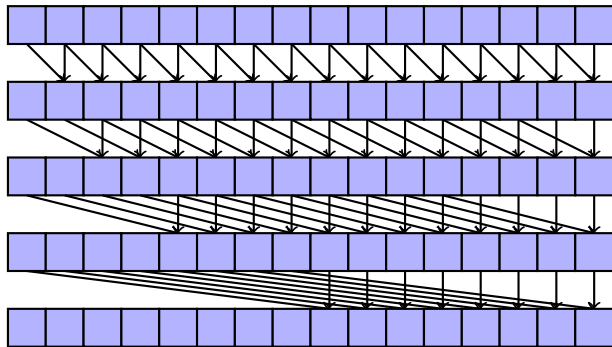
where N is the input size. (Think: N large, $f(x, y) = x + y$)

- ▶ Prefix Sum/Cumulative Sum
- ▶ Abstract view of: loop-carried dependence
- ▶ Also possible: Segmented Scan

Scan: Graph



Scan: Implementation



Work-efficient?

Scan: Implementation II

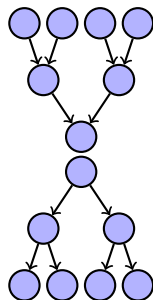
Two sweeps: Upward, downward, both tree-shape

On upward sweep:

- ▶ Get values L and R from left and right child
- ▶ Save L in local variable $Mine$
- ▶ Compute $Tmp = L + R$ and pass to parent

On downward sweep:

- ▶ Get value Tmp from parent
- ▶ Send Tmp to left child
- ▶ Send $Tmp + Mine$ to right child



Scan: Examples

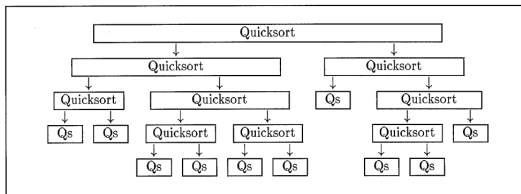
Name examples of Prefix Sums/Scans:



Data-parallel language: Goals

Goal: Design a full data-parallel programming language

Example: What should the (asymptotic) execution time for Quicksort be?



Question: What parallel primitive could be used to realize this?

NESL Example: String Search

```
teststr = "string strap asop string" : [char]
>>> candidates = [0:#teststr-5];
candidates = [0, 1, 2, 3, .... : [int]
>>> {a == 's: a in teststr -> candidates};
it = [T, F, F, F, F, F, F, T, F, F....] : [bool]
>>> candidates = {c in candidates;
...      a in teststr -> candidates | a == 's};
candidates = [0, 7, 13, 20, 24] : [int]
>>> candidates = {c in candidates;
...      a in teststr -> {candidates+1:candidates}
...      | a == 't};
```

- ▶ Work and depth of this example?
- ▶ NESL specifies work and depth for its constructs
- ▶ How can scans be used to realize this?

[Blelloch '95](#)

Array Languages

Idea:

- ▶ Operate on entire array at once
- ▶ Inherently data-parallel

Examples:

- ▶ APL, numpy
- ▶ Tensorflow (talk on Friday), Pytorch

Important axes of distinction:

- ▶ Lazy or eager
- ▶ Imperative (with in-place modification) or pure/functional

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Polyhedral Model: What?

Outline

Introduction

Machine Abstractions

Performance: Expectation, Experiment, Observation

Performance-Oriented Languages and Abstractions

Polyhedral Representation and Transformation

Polyhedral Model: What?

Basic Object: Presburger Set

Think of the **problem statement** here as representing an arbitrary-size (e.g.: dependency) graph.

Presburger sets correspond to a subset of predicate logic acting on tuples of integers.

Important: Think of this as a mathematical tool that can be used in many settings.

Basic Object: Presburger Set

Terms:

- ▶ Variables, Integer Constants
- ▶ $+$, $-$
- ▶ $\lfloor \cdot / d \rfloor$

Predicates:

- ▶ $(\text{Term}) \leq (\text{Term})$
- ▶ $(\text{Pred}) \wedge (\text{Pred}), (\text{Pred}) \vee (\text{Pred}), \neg(\text{Pred})$
- ▶ $\exists v : (\text{Pred})(v)$

Sets: integer tuples for which a predicate is true

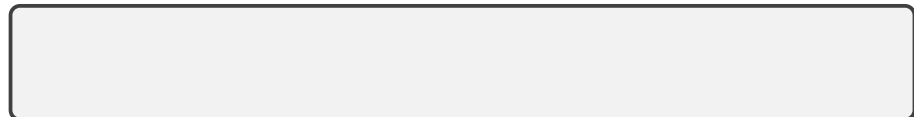
[Verdoolaege '13](#)

Presburger Sets: Reasoning

What's "missing"? Why?

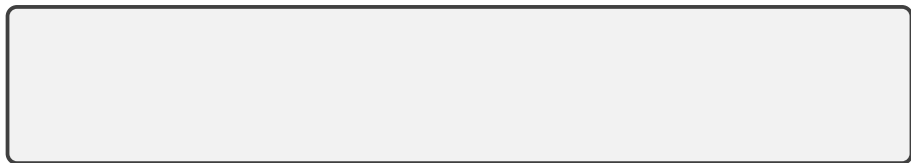
A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide reasoning for the first question.

Why is this called 'quasi-affine'?

A large, empty rectangular box with a thin black border and rounded corners, intended for the user to provide reasoning for the second question.

Presburger Sets: Reasoning

What do the resulting sets have to do with polyhedra? When are they convex?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

Why polyhedra? Why not just rectangles?

A large, empty rectangular box with a thin black border and rounded corners, intended for a user to provide an answer to the question above.

Demo: Constructing and Operating on Presburger Sets

Demo: lang/Operating on Presburger Sets

Making Use of Presburger Sets

- ▶ Loop Domains
- ▶ Array Access Relations (e.g. write, read: per statement)
- ▶ Schedules, with “lexicographic time”
- ▶ Dependency graphs
- ▶ (E.g. cache) interference graphs


Q: Specify domain and range for the relations above.

Example: Dependency Graph

Given:

- ▶ Write access relation W : Loop domain \rightarrow array indices
- ▶ Read access relation R
- ▶ Schedule S for statement S_i : Loop domain $D \rightarrow$ lex. time of statement instance
- ▶ Relation \prec : Lexicographic 'before'

Find the dependency graph:



[Verdoolaege '13](#)

Example: Last Instance

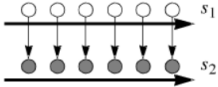
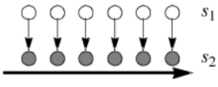
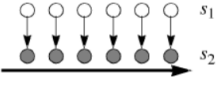
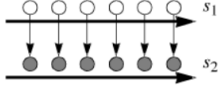
Given:

- ▶ Write access relation W : Loop domain \rightarrow array indices
- ▶ Read access relation R
- ▶ Schedule S for statement S_i : Loop domain $D \rightarrow$ lex. time of statement instance
- ▶ Relation \prec : Lexicographic 'before'

Find the statement instances accessing array element:

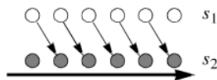
Find the *last* statement instance accessing array element:

Primitive Transformations (I)

SOURCE CODE	PARTITION	TRANSFORMED CODE
<pre> for (i=1; i<=N; i++) Y[i] = Z[i]; /*s1*/ for (j=1; j<=N; j++) X[j] = Y[j]; /*s2*/ </pre> 	<p>Fusion</p> $s_1 : p = i$ $s_2 : p = j$	<pre> for (p=1; p<=N; p++){ Y[p] = Z[p]; X[p] = Y[p]; } </pre> 
<pre> for (p=1; p<=N; p++){ Y[p] = Z[p]; X[p] = Y[p]; } </pre> 	<p>Fission</p> $s_1 : i = p$ $s_2 : j = p$	<pre> for (i=1; i<=N; i++) Y[i] = Z[i]; /*s1*/ for (j=1; j<=N; j++) X[j] = Y[j]; /*s2*/ </pre> 

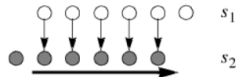
Primitive Transformations (II)

```
for (i=1; i<=N; i++) {
    Y[i] = Z[i]; /*s1*/
    X[i] = Y[i-1]; /*s2*/
}
```

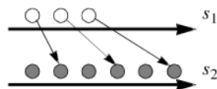


Re-indexing
 $s_1 : p = i$
 $s_2 : p = i - 1$

```
if (N>=1) X[1]=Y[0];
for (p=1; p<=N-1; p++){
    Y[p]=Z[p];
    X[p+1]=Y[p];
}
if (N>=1) Y[N]=Z[N];
```

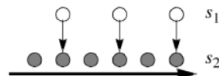


```
for (i=1; i<=N; i++)
    Y[2*i] = Z[2*i]; /*s1*/
for (j=1; j<=2N; j++)
    X[j]=Y[j]; /*s2*/
```

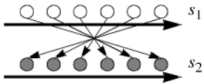



Scaling
 $s_1 : p = 2 * i$
 $(s_2 : p = j)$

```
for (p=1; p<=2*N; p++){
    if (p mod 2 == 0)
        Y[p] = Z[p];
    X[p] = Y[p];
}
```



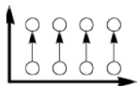
Primitive Transformations (III)

SOURCE CODE	PARTITION	TRANSFORMED CODE
<pre> for (i=0; i<=N; i++) Y[N-i] = Z[i]; /*s1*/ for (j=0; j<=N; j++) X[j] = Y[j]; /*s2*/ </pre> 	<p>Reversal</p> $s_1 : p = N - i$ $(s_2 : p = j)$	<pre> for (p=0; p<=N; p++){ Y[p] = Z[N-p]; X[p] = Y[p]; } </pre> 

[Aho/Ullman/Sethi '07]

Primitive Transformations (IV)

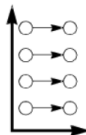
```
for (i=1; i<=N; i++)
  for (j=0; j<=M; j++)
    Z[i,j] =
      Z[i-1,j];
```



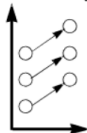
Permutation

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} i \\ j \end{bmatrix}$$

```
for (p=0; p<=M; p++)
  for (q=1; q<=N; q++)
    Z[q,p] = Z[q-1,p]
```



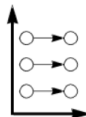
```
for (i=1; i<=N+M-1; i++)
  for (j=max(1,i+N);
       j<=min(i,M); j++)
    Z[i,j] =
      Z[i-1,j-1];
```



Skewing

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

```
for (p=1; p<=N; p++)
  for (q=1; q<=M; q++)
    Z[p,q-p] =
      Z[p-1,q-p-1]
```



Example: Last Instance

Given:

- ▶ Dependency relation D

Check whether a transformed schedule S' is valid:

A peek under the hood: Fourier-Motzkin Elimination

INPUT: A polyhedron S with variables x_1, \dots, x_n

OUTPUT: x_1, \dots, x_{n-1}

- ▶ Let C be all the constraints in S involving x_n .
- ▶ For every pair of a lower and an upper bound on x_n in C :

$$\begin{aligned} L &\leq c_1 x_n, \\ &\leq c_2 x_n \leq U, \end{aligned}$$

create the new constraint $c_2 L \leq c_1 U$.

- ▶ Divide by the GCD of c_1, c_2 if applicable.
- ▶ If the new constraint is not satisfiable, S was empty.
- ▶ Let $S' = S \setminus C \cup \bigcup_{L,U} \{c_2 L \leq c_1 U\}$.

[Aho/Ullman/Sethi '07]

Q: How can this help implement an emptiness check?