GPU Microarchitecture Note Set 1b—Cores Intro

The material in this set is a summary. Set 2 will cover cores in more detail.

Topics in This Set

- Core Definition
- Capability and Performance Measures
- Heavy Cores v. Light Cores

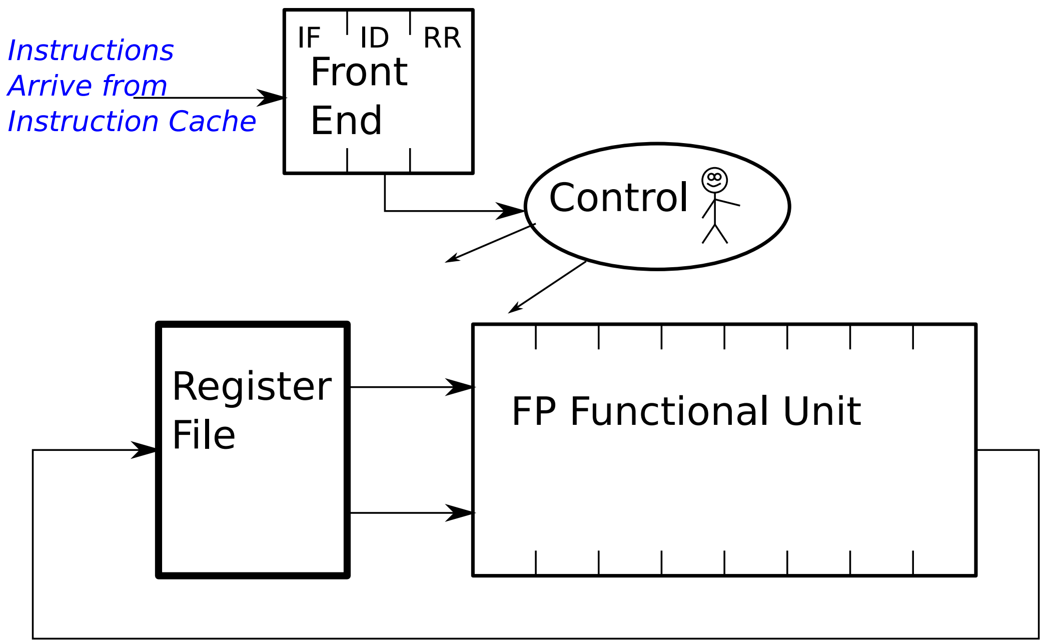
Core:

Hardware needed to execute a thread.

Sometimes called a CPU (central processing unit).

Each core has:

- Hardware to fetch instructions.
- *Functional units* to perform arithmetic operations.
- *Register files* to hold intermediate (working, temporary) data values.
- Hardware to decode and orchestrate instruction execution.



Measures of Capability and Performance of a Core

Consider the following program:

```
i1 :  add.s r1,  r2,  r3   # Note: r2 = 300,  r3 = 3     Computes r1 = r2 + r3
i2 :  add.s r4,  r5,  r6   # Note: r5 = 400,  r6 = 4     Computes r4 = r5 + r6
i3 :  add.s r7,  r8,  r9
i4 :  add.s r10, r11, r12
..
i999:  add.s r25, r26, r27
i1000: add.s r28, r29, r30  # Note: r29 = 1200,  r30 = 12
```

It consists of 1000 instructions and 1000 FP operations.

Measures of Capability and Performance of a Core

Instruction Bandwidth (IB):
Peak rate at which core can execute instructions.

A measure of what a core is *capable* of.

Usually measured in *instructions per cycle*, abbreviated *IPC*.

To obtain *instructions per second* multiply IPC by clock frequency.

A *w-way superscalar processor* has an instruction bandwidth of w.

Instruction Bandwidth of Real Cores

- Early RISC processors: 1 insn/cyc per core.

- POWER 7 (IBM): 6 insn/cyc per core.

- Intel Core: 4 insn/cyc per core.

Instruction Throughput:
Instruction execution rate achieved by some program on some core.

A measure of the performance of some program on some core.

Also measured in IPC.

The throughput cannot be higher than the bandwidth.

Understanding Throughput

Higher throughput means core hardware used more efficiently.

A program with higher throughput is not necessarily faster . . .
. . . because execution time depends on number of instructions.

Saturation:
A situation in which a performance measure matches the corresponding capability.

Example of Saturation

Program P is said to *saturate* core C . . .
. . . if core C has an instruction bandwidth of w . . .
. . . and program P runs with an instruction throughput of w on C.

Dynamic Instruction Count:

The number of instructions executed by a program.

Example— Usage of Terms

A core has an IB of $4\,\mathrm{insn/cyc}$*, and a clock frequency of* $1\,\mathrm{GHz}$*. A program has a dynamic instruction count of* 2×10^9 *instructions and takes 1 second to run on the core. Are we happy?*

Instruction throughput is: $\frac{2\times10^9\,\mathrm{insn}}{10^9\,\mathrm{cyc}} = 2\,\mathrm{insn/cyc}$.

That's half of the core's potential rate.

Let's assume the code was challenging …
… so achieving half the peak is an accomplishment …
… so we are happy.

Other Core Capabilities

Here are other core capabilities of interest to us:

- Single Precision Floating Point Rate, measured in FLOPS.

- Double Precision Floating Point Rate, measured in FLOPS.

- Off-Chip Data Bandwidth, measured in bytes per second.

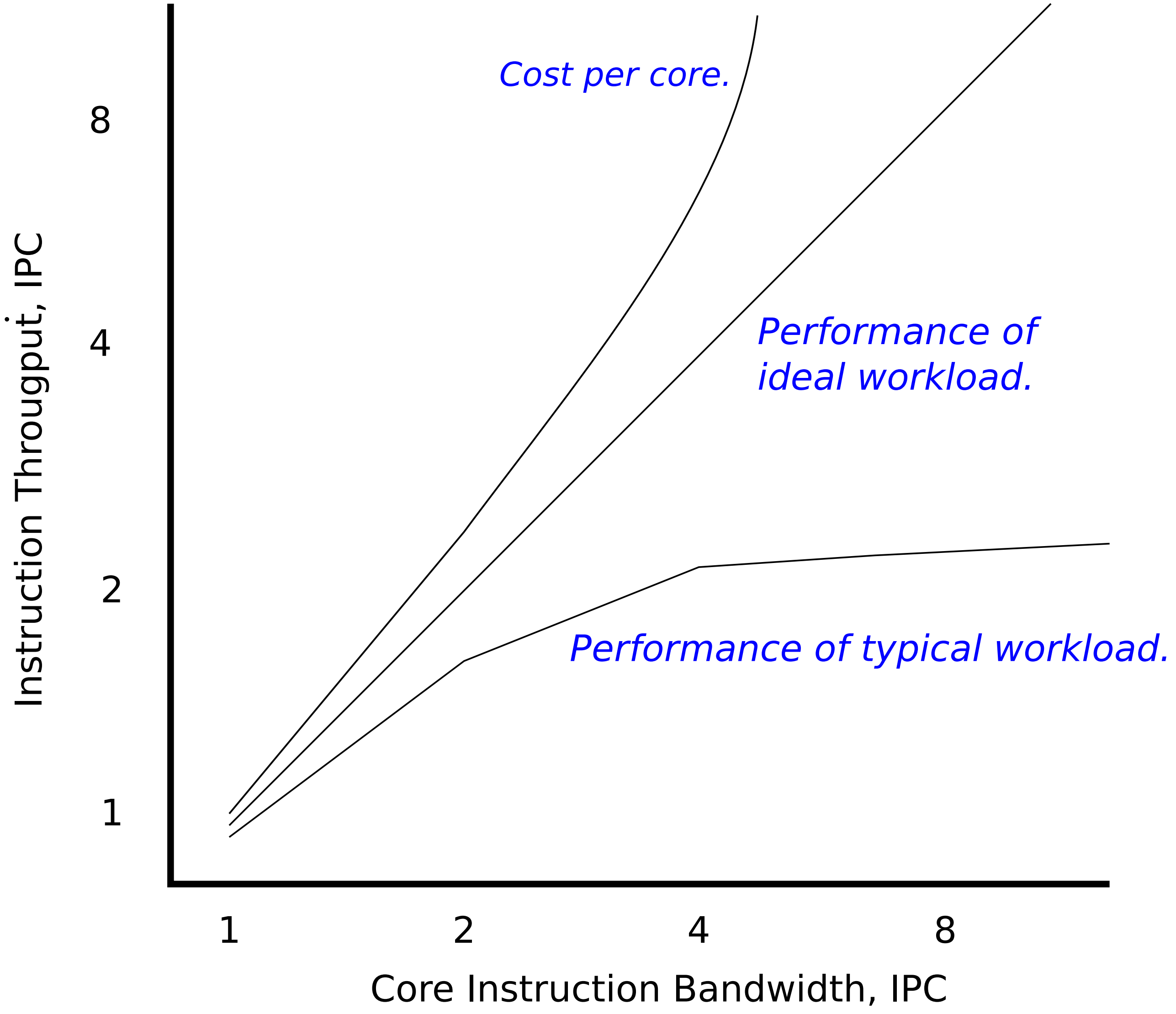
For each there is a corresponding performance number.

Why not make instruction bandwidth large?

An 8-way core is only slightly better …
… than 2 1-way cores …
… but costs 4 times as much!

For almost linear portion of curve, bigger core is an easy choice.

Distance between the ideal and typical curve…
… is the price of avoiding parallel programming.



Implications

Parallel programming can't be avoided.

Why not make IB small, say 1, and have lots of cores?

Parallel programming is hard.

Parallel programming is another thing you have to do.

Speedup may not be linear.

Heavy Weight Core:
A core designed to execute a single thread quickly.

Heavy weight cores have large area and high power consumption.

Energy per instruction is high.

General-purpose CPUs, such as those found in home computers, consist of heavy weight cores.

Light Weight Core:
A core designed for efficiency.

Light weight cores have small area.

Energy per instruction is low.

Multiple Core Chips

Multi-Core Chip:
A chip with a few heavy-weight cores.

Many-Core Chip:
A chip with many light-weight cores.

Comparison

Many-core chip has higher instruction bandwidth (counting all cores).

Multi-core chip has higher instruction bandwidth (counting one core).

Execution of Multithreaded Programs

Consider a system with c cores and a program with r threads.

Typically the OS will distribute the r threads evenly over the c cores.

If $r < c$ then $c - r$ cores will sit idle.

If $r > c$ then a core may have more than one thread assigned.