

Programming Modern Processors (2) && Parallel Architectures

Hung-Wei Tseng

Recap: addressing hazards

- Structural hazards
 - Stall
 - Modify hardware design
- Control hazards
 - Stall
 - Static prediction
 - Dynamic prediction
- Data Hazards
 - Stall
 - Data forwarding
 - Dynamic instruction scheduling

AMD Zen 3 (RyZen 5000 Series)

3-issue memory pipeline

4-issue integer pipeline + 1 additional branch

$$MinCPI = \frac{1}{8}$$

$$MinINTInst . CPI = \frac{1}{4}$$

$$MinMEMInst . CPI = \frac{1}{3}$$

$$MinBRIInst . CPI = \frac{1}{2}$$

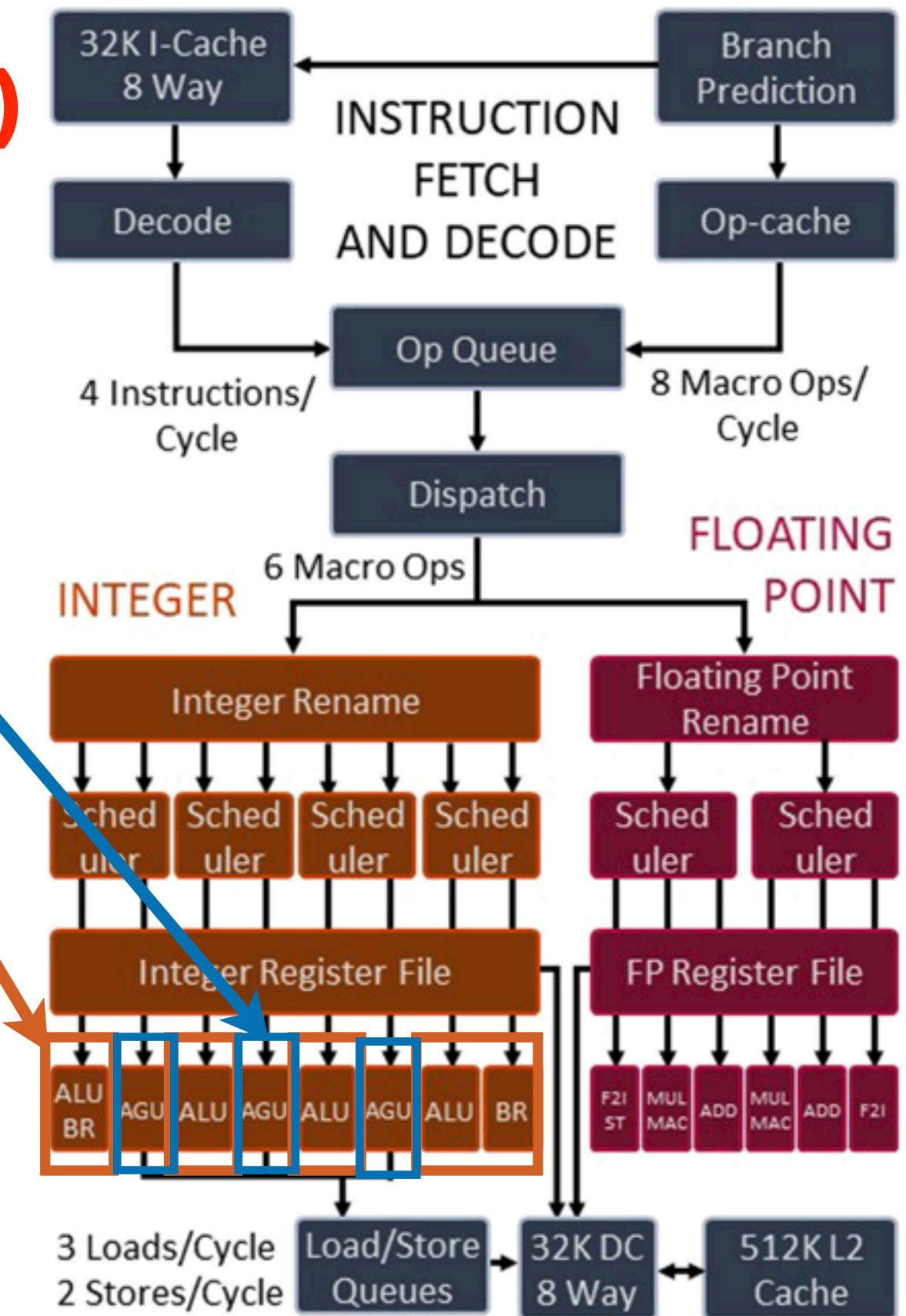


FIGURE 1. “Zen 3” block diagram.

Recap: Intel Skylake

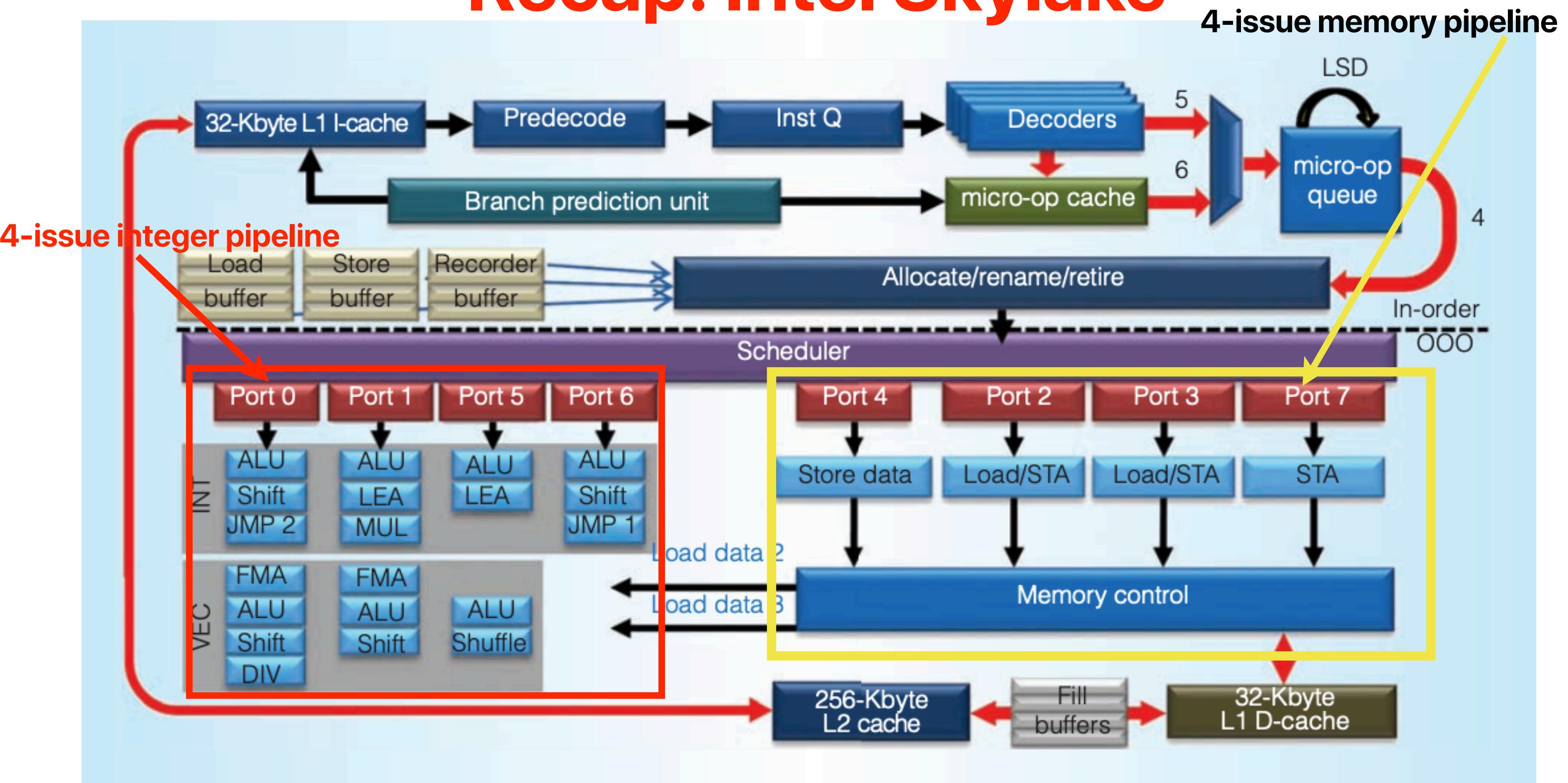
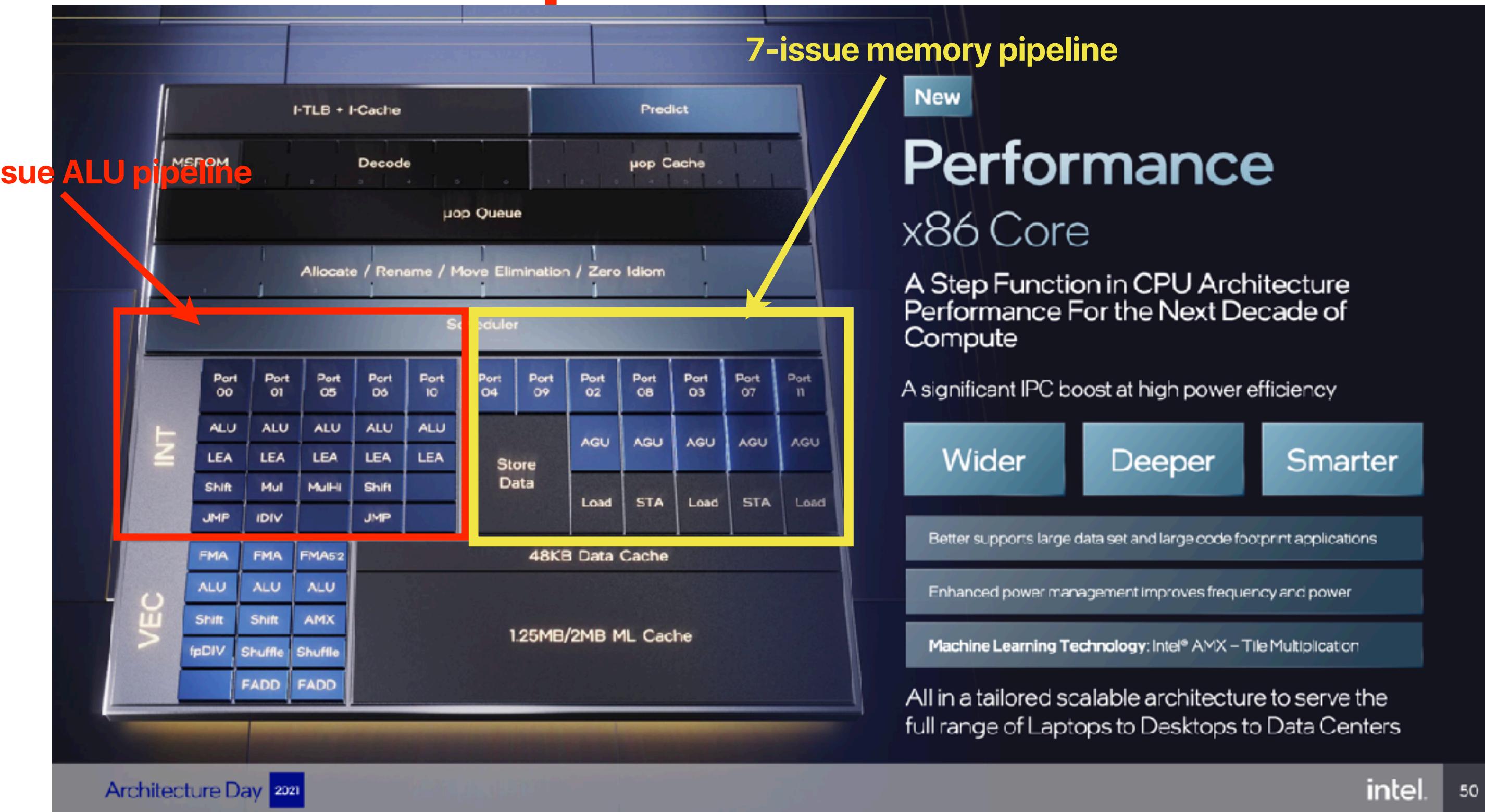


Figure 4. Skylake core block diagram.

Recap: Intel Alder Lake



Summary: Characteristics of modern processor architectures

- SuperScalar
 - Multiple-issue pipelines with multiple functional units available
 - Multiple instructions can execute in the same pipeline
- Register renaming + Reorder Buffer (ROB) — enable dynamic OoO scheduling to issue/execute instructions whenever possible
- Cache
- Branch predictors

Five implementations

- Which of the following implementations will perform the best on modern pipeline processors?

A

```
inline int __popcount(uint64_t x){  
    int c=0;  
    while(x) {  
        c += x & 1;  
        x = x >> 1;  
    }  
    return c;  
}
```

B

```
inline int __popcount(uint64_t x){  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    while(x) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```


C

B

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    while(x) {  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
    }  
    return c;  
}
```


D

D

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    for (uint64_t i = 0; i < 16; i++) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

E

E

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    for (uint64_t i = 0; i < 16; i++) {  
        switch((x & 0xF)) {  
            case 1: c+=1; break;  
            case 2: c+=1; break;  
            case 3: c+=2; break;  
            case 4: c+=1; break;  
            case 5: c+=2; break;  
            case 6: c+=2; break;  
            case 7: c+=3; break;  
            case 8: c+=1; break;  
            case 9: c+=2; break;  
            case 10: c+=2; break;  
            case 11: c+=3; break;  
            case 12: c+=2; break;  
            case 13: c+=3; break;  
            case 14: c+=3; break;  
            case 15: c+=4; break;  
            default: break;  
        }  
        x = x >> 4;  
    }  
    return c;  
}
```

Outline

- Programming on Modern Processors (cont.)
- Multithreaded architectures

Why is B better than A?

- How many of the following statements explains the reason why B outperforms A with compiler optimizations
 - B has lower dynamic instruction count than A
 - B has significantly lower branch mis-prediction rate than A
 - B has significantly fewer branch instructions than A
 - B can incur fewer data hazards

- A. 0
- B. 1
- C. 2
- D. 3
- E. 4

A

```
inline int __popcount(uint64_t x) {
    int c=0;
    while(x) {
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

B

```
inline int __popcount(uint64_t x) {
    int c = 0;
    while(x) {
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

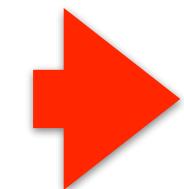
Why is B better than A?

A

```
inline int __popcount(uint64_t x){  
    int c=0;  
    while(x) {  
        c += x & 1;  
        x = x >> 1;  
    }  
    return c;  
}
```

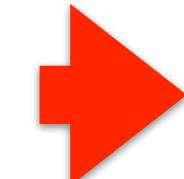
B

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    while(x) {  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
        c += x & 1;  
        x = x >> 1;  
    }  
    return c;  
}
```



5*n instructions

```
movl %eax, %ecx  
andl $1, %ecx  
addl %ecx, %edx  
shrq %rax  
jne .L6
```



$15*(n/4) = 3.75*n$ instructions

Only one branch for four iterations in A

**compiler reorder
instructions and rename
registers to mitigate
hazards and exploit ILP**

```
%ecx, %eax  
$1, %eax  
%edx, %eax  
%rcx, %rdx  
%rdx  
$1, %edx  
%eax, %edx  
%rcx, %rax  
$2, %rax  
$1, %eax  
%edx, %eax  
%rcx, %rdx  
$3, %rdx  
$1, %edx  
%eax, %edx  
$4, %rcx  
.L6
```

Why is B better than A?

- How many of the following statements explains the reason why B outperforms A with compiler optimizations

- ① B has lower dynamic instruction count than A
- ② B has significantly lower branch mis-prediction rate than A
- ③ B has significantly fewer branch instructions than A
- ④ B can incur fewer data hazards

A. 0

B. 1

C. 2

D. 3

E. 4

A

```
inline int __popcount(uint64_t x) {
    int c=0;
    while(x) {
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

B

```
inline int __popcount(uint64_t x) {
    int c = 0;
    while(x) {
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

Why is C better than B?

- How many of the following statements explains the reason why B outperforms C with compiler optimizations
 - C has lower dynamic instruction count than B
 - C has significantly lower branch mis-prediction rate than B
 - C has significantly fewer branch instructions than B
 - C can incur fewer data hazards

A. 0

B. 1

C. 2

D. 3

E. 4



```
inline int __popcount(uint64_t x) {
    int c = 0;
    int table[16] = {0, 1, 1, 2, 1,
                     2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};
    while(x) {
        c += table[(x & 0xF)];
        x = x >> 4;
    }
    return c;
}
```



```
inline int __popcount(uint64_t x) {
    int c = 0;
    while(x) {
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

Why is C better than B?

- How many of the following statements explains the reason why B outperforms C with compiler optimizations
- ① C has lower dynamic instruction count than B
—C only needs one load, one add, one shift, the same amount of iterations
 - ② C has significantly lower branch mis-prediction rate than B
—the same number being predicted.
 - ③ C has significantly fewer branch instructions than B —the same amount of branches
 - ④ C can incur fewer data hazards
—Probably not. In fact, the load may have negative effect without architectural supports

A. 0

B. 1

C. 2

D. 3

E. 4

```
inline int __popcount(uint64_t x) {
    int c = 0;
    int table[16] = {0, 1, 1, 2, 1,
                     2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};
    while(x) {
        c += table[(x & 0xF)];
        x = x >> 4;
    }
    return c;
}
```

```
inline int __popcount(uint64_t x) {
    int c = 0;
    while(x) {
        c += x & 1;
        x = x >> 1;
    }
    return c;
}
```

Why is D better than C?

- How many of the following statements explains the main reason why B outperforms C with compiler optimizations
 - D has lower dynamic instruction count than C
 - D has significantly lower branch mis-prediction rate than C
 - D has significantly fewer branch instructions than C
 - D can incur fewer data hazards than C

A. 0

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    while(x) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

B. 1

C. 2

D. 3

E. 4



```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    for (uint64_t i = 0; i < 16; i++) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

Why is D better than C?

- How many of the following statements explains the main reason why B outperforms C with compiler optimizations

① D has lower dynamic instruction count than C

— Compiler can do loop unrolling — no branches

② D has significantly lower branch mis-prediction rate than C

— Could be

③ D has significantly fewer branch instructions than C

— maybe eliminated through loop unrolling...

④ D can incur fewer data hazards than C

— about the same

A. 0

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    while(x) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

B. 1

C. 2

D. 3

E. 4

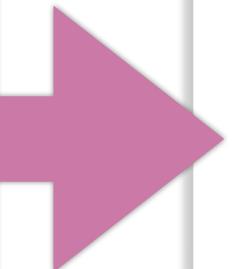
C

D

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    for (uint64_t i = 0; i < 16; i++) {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

Loop unrolling eliminates all branches!

```
inline int __popcount(uint64_t x) {
    int c = 0;
    int table[16] = {0, 1, 1, 2, 1,
2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};
    for (uint64_t i = 0; i < 16; i++)
    {
        c += table[(x & 0xF)];
        x = x >> 4;
    }
    return c;
}
```



Why is E the slowest?

- How many of the following statements explains the main reason why B outperforms C with compiler optimizations
 - E has the most dynamic instruction count
 - E has the highest branch mis-prediction rate
 - E has the most branch instructions
 - E can incur the most data hazards than others

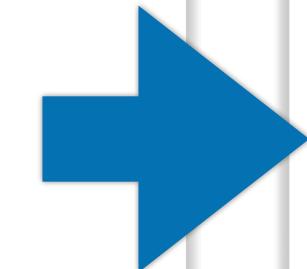
- A. 0
- B. 1
- C. 2
- D. 3
- E. 4

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    for (uint64_t i = 0; i < 16; i++)  
    {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    for (uint64_t i = 0; i < 16; i++)  
    {  
        switch((x & 0xF))  
        {  
            case 1: c+=1; break;  
            case 2: c+=1; break;  
            case 3: c+=2; break;  
            case 4: c+=1; break;  
            case 5: c+=2; break;  
            case 6: c+=2; break;  
            case 7: c+=3; break;  
            case 8: c+=1; break;  
            case 9: c+=2; break;  
            case 10: c+=2; break;  
            case 11: c+=3; break;  
            case 12: c+=2; break;  
            case 13: c+=3; break;  
            case 14: c+=3; break;  
            case 15: c+=4; break;  
            default: break;  
        }  
        x = x >> 4;  
    }  
    return c;  
}
```

Why is E the slowest?

```
inline int __popcount(uint64_t x) {
    int c = 0;
    for (uint64_t i = 0; i < 16; i++)
    {
        switch((x & 0xF))
        {
            case 1: c+=1; break;
            case 2: c+=1; break;
            case 3: c+=2; break;
            case 4: c+=1; break;
            case 5: c+=2; break;
            case 6: c+=2; break;
            case 7: c+=3; break;
            case 8: c+=1; break;
            case 9: c+=2; break;
            case 10: c+=2; break;
            case 11: c+=3; break;
            case 12: c+=2; break;
            case 13: c+=3; break;
            case 14: c+=3; break;
            case 15: c+=4; break;
            default: break;
        }
        x = x >> 4;
    }
    return c;
}
```



.L11:

```
    movq    %r9, %rcx
    andl    $15, %ecx
    movslq  (%r8,%rcx,4), %rcx
    addq    %r8, %rcx
    notrack jmp     *%rcx
```

.L7:

```
    .long   .L5-.L7
    .long   .L10-.L7
    .long   .L10-.L7
    .long   .L9-.L7
    .long   .L10-.L7
    .long   .L9-.L7
    .long   .L9-.L7
    .long   .L8-.L7
    .long   .L10-.L7
    .long   .L9-.L7
    .long   .L9-.L7
    .long   .L8-.L7
    .long   .L9-.L7
    .long   .L8-.L7
    .long   .L9-.L7
    .long   .L8-.L7
    .long   .L6-.L7
```

.L8:

```
    addl    $3, %eax
```

.L5:

```
    shrq    $4, %r9
    subq    $1, %rsi
    jne     .L11
    cltq
    addq    %rax, %rbx
    subl    $1, %edi
    jne     .L12
```

.L9:

```
.cfi_restore_state
    addl    $2, %eax
    jmp    .L5
    .p2align 4,,10
    .p2align 3
```

.L10:

```
    addl    $1, %eax
    jmp    .L5
    .p2align 4,,10
    .p2align 3
```

.L6:

```
    addl    $4, %eax
    jmp    .L5
```

Why is E the slowest?

- How many of the following statements explains the main reason why B outperforms C with compiler optimizations

- ① E has the most dynamic instruction count
- ② E has the highest branch mis-prediction rate
- ③ E has the most branch instructions
- ④ E can incur the most data hazards than others

- A. 0
- B. 1
- C. 2
- D. 3
- E. 4

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    int table[16] = {0, 1, 1, 2, 1,  
                    2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4};  
    for (uint64_t i = 0; i < 16; i++)  
    {  
        c += table[(x & 0xF)];  
        x = x >> 4;  
    }  
    return c;  
}
```

```
inline int __popcount(uint64_t x) {  
    int c = 0;  
    for (uint64_t i = 0; i < 16; i++)  
    {  
        switch((x & 0xF))  
        {  
            case 1: c+=1; break;  
            case 2: c+=1; break;  
            case 3: c+=2; break;  
            case 4: c+=1; break;  
            case 5: c+=2; break;  
            case 6: c+=2; break;  
            case 7: c+=3; break;  
            case 8: c+=1; break;  
            case 9: c+=2; break;  
            case 10: c+=2; break;  
            case 11: c+=3; break;  
            case 12: c+=2; break;  
            case 13: c+=3; break;  
            case 14: c+=3; break;  
            case 15: c+=4; break;  
            default: break;  
        }  
        x = x >> 4;  
    }  
    return c;  
}
```

Hardware acceleration

- Because `popcount` is important, both intel and AMD added a `POPCNT` instruction in their processors with SSE4.2 and SSE4a
- In C/C++, you may use the intrinsic “`_mm_popcnt_u64`” to get # of “1”s in an unsigned 64-bit number
 - You need to compile the program with `-m64 -msse4.2` flags to enable these new features

```
#include <smmintrin.h>
inline int popcount(uint64_t x) {
    int c = _mm_popcnt_u64(x);
    return c;
}
```

How're you going to optimize the code?

```
for(i=0;i<size;i++)  {  
    if (a[i] < y && a[i] > x)  
        sum++;  
}
```

Ideas?

```
for(i=0;i<size;i++) {  
    if (a[i] < y && a[i] > x)  
        sum++;  
}
```

```
for(i=0;i<size;i++) {  
    sum_2 = sum+1;  
    sum = (a[i] < y && a[i] > x) ? sum_2: sum;  
}
```

```
for(i=0;i<size;i++) {  
    if (a[i] < y & a[i] > x)  
        sum++;  
}
```

```
for(i=0;i<size;i++) {  
    sum+=(a[i]<y && a [i]>x);  
}
```

**Refresh our minds! What are the tips
you have in mind in writing efficient
programs on modern processors?**

Tips of programming on modern processors

- Minimize the critical path operations
 - Don't forget about optimizing cache/memory locality first!
 - Memory latencies are still way longer than any arithmetic instruction
 - Can we use arrays/hash tables instead of lists?
 - Branch can be expensive as pipeline get deeper
 - Sorting
 - Loop unrolling
 - Still need to carefully avoid long latency operations (e.g., mod)
- Since processors have multiple functional units — code must be able to exploit instruction-level parallelism
 - Hide as many instructions as possible under the “critical path”
 - Try to use as many different functional units simultaneously as possible
- Modern processors also have accelerated instructions

Architecture:	x86_64
CPU op-mode(s):	32-bit, 64-bit
Byte Order:	Little Endian
Address sizes:	48 bits physical, 48 bits virtual
CPU(s):	16
On-line CPU(s) list:	0-15
Thread(s) per core:	2
Core(s) per socket:	8
Socket(s):	1
NUMA node(s):	1
Vendor ID:	AuthenticAMD
CPU family:	25
Model:	80
Model name:	AMD Ryzen 7 5700G with Radeon Graphics
Stepping:	0

Demo: ILP within a program

- perf is a tool that captures performance counters of your processors and can generate results like branch mis-prediction rate, cache miss rates and ILP.

Wider-issue processors won't give you much more

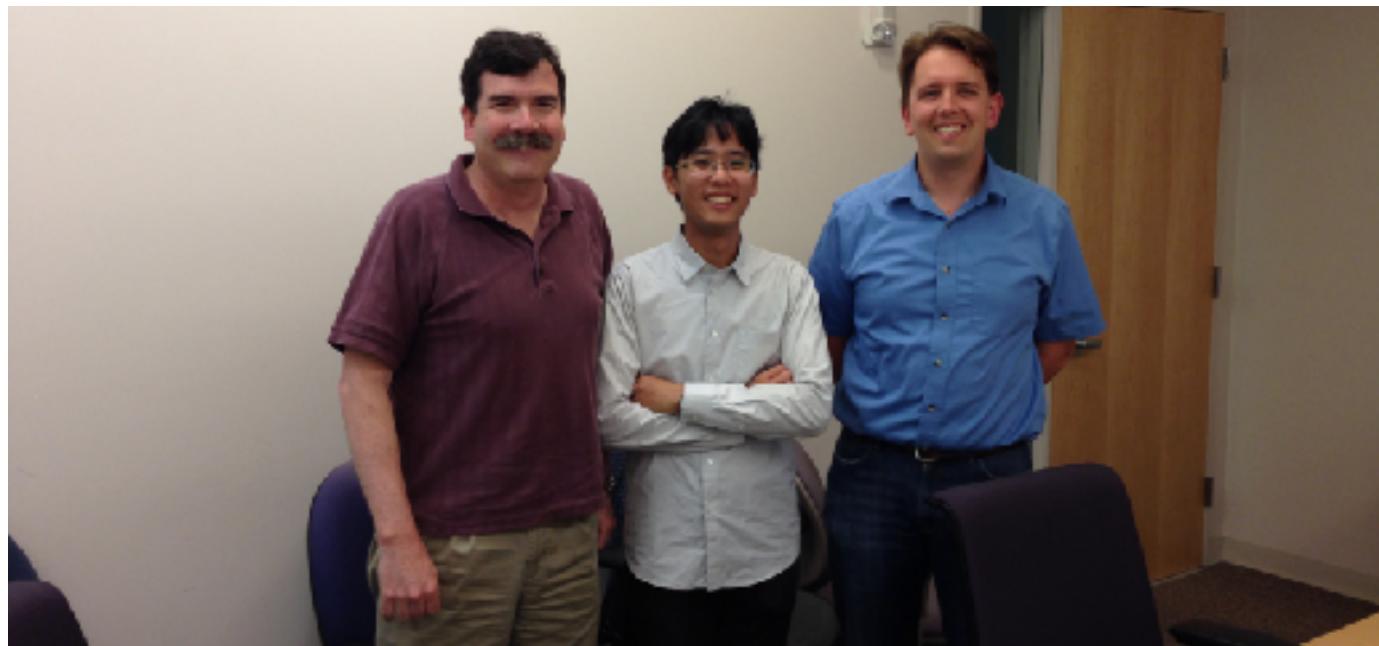
Program	IPC	BP Rate %	I cache %MPCI	D cache %MPCI	L2 cache %MPCI
compress	0.9	85.9	0.0	3.5	1.0
eqntott	1.3	79.8	0.0	0.8	0.7
m88ksim	1.4	91.7	2.2	0.4	0.0
MPsim	0.8	78.7	5.1	2.3	2.3
applu	0.9	79.2	0.0	2.0	1.7
apsi	0.6	95.1	1.0	4.1	2.1
swim	0.9	99.7	0.0	1.2	1.2
tomcatv	0.8	99.6	0.0	7.7	2.2
pmake	1.0	86.2	2.3	2.1	0.4

Program	IPC	BP Rate %	I cache %MPCI	D cache %MPCI	L2 cache %MPCI
compress	1.2	86.4	0.0	3.9	1.1
eqntott	1.8	80.0	0.0	1.1	1.1
m88ksim	2.3	92.6	0.1	0.0	0.0
MPsim	1.2	81.6	3.4	1.7	2.3
applu	1.7	79.7	0.0	2.8	2.8
apsi	1.2	95.6	0.2	3.1	2.6
swim	2.2	99.8	0.0	2.3	2.5
tomcatv	1.3	99.7	0.0	4.2	4.3
pmake	1.4	82.7	0.7	1.0	0.6

Table 5. Performance of a single 2-issue superscalar processor.

Table 6. Performance of the 6-issue superscalar processor.

Simultaneous multithreading

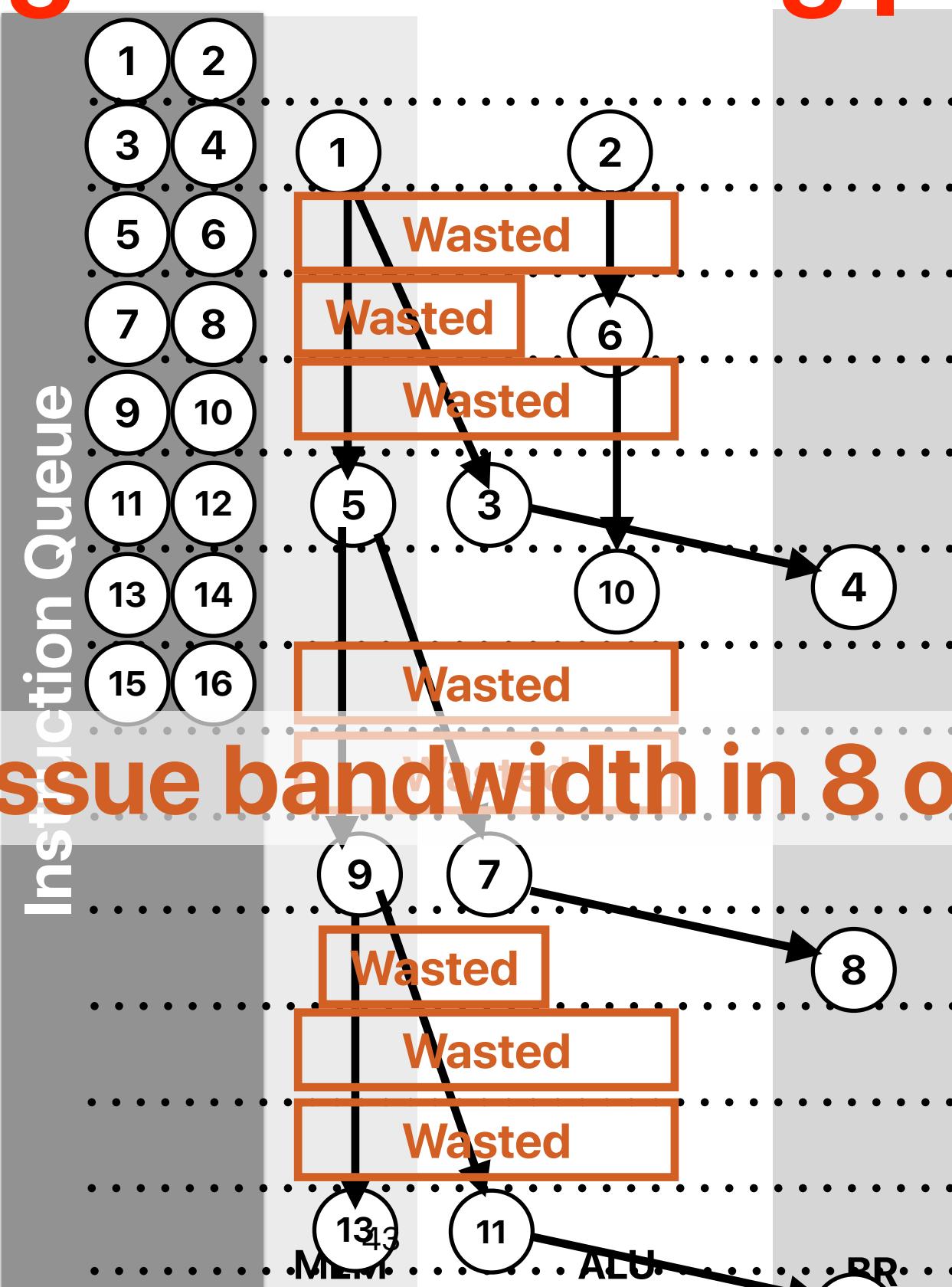


Simultaneous multithreading

- Invented by Dean Tullsen (Now a professor at UCSD CSE)
- The processor can schedule instructions from different threads/processes/programs
- Fetch instructions from different threads/processes to fill the not utilized part of pipeline
 - Exploit “thread level parallelism” (TLP) to solve the problem of insufficient ILP in a single thread
 - You need to create an illusion of multiple processors for OSs

2-issue register renaming pipeline

```
① movq    8(%rdi), %rdi  
② addl    $1, %eax  
③ testq   %rdi, %rdi  
④ jne     .L3  
⑤ movq    8(%rdi), %rdi  
⑥ addl    $1, %eax  
⑦ testq   %rdi, %rdi  
⑧ jne     .L3  
⑨ movq    8(%rdi), %rdi  
⑩ addl    $1, %eax  
⑪ testq   %rdi, %rdi  
⑫ jne     .L3
```

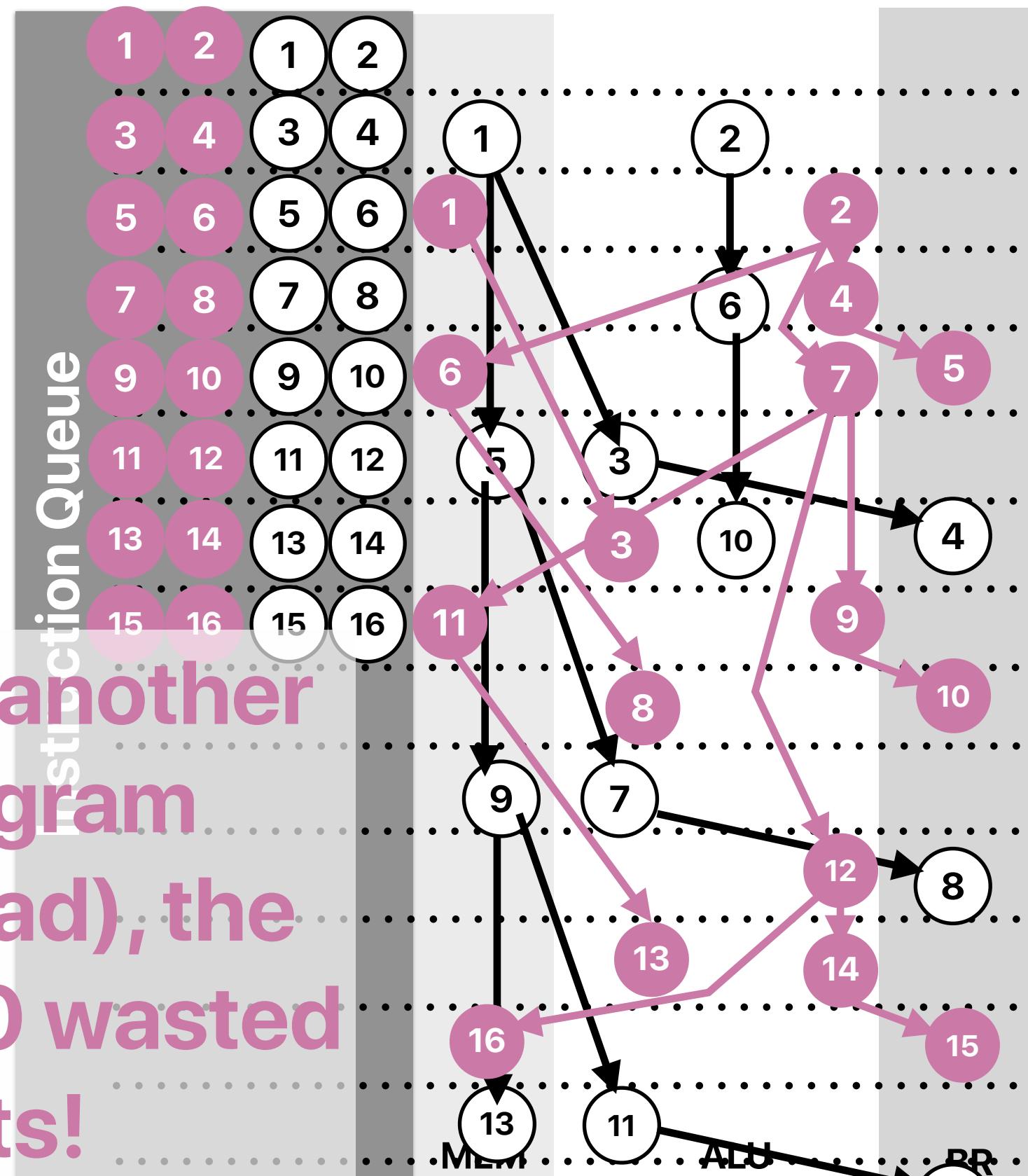


We're wasting the issue bandwidth in 8 out of 12 cycles

Concept: Simultaneous Multithreading

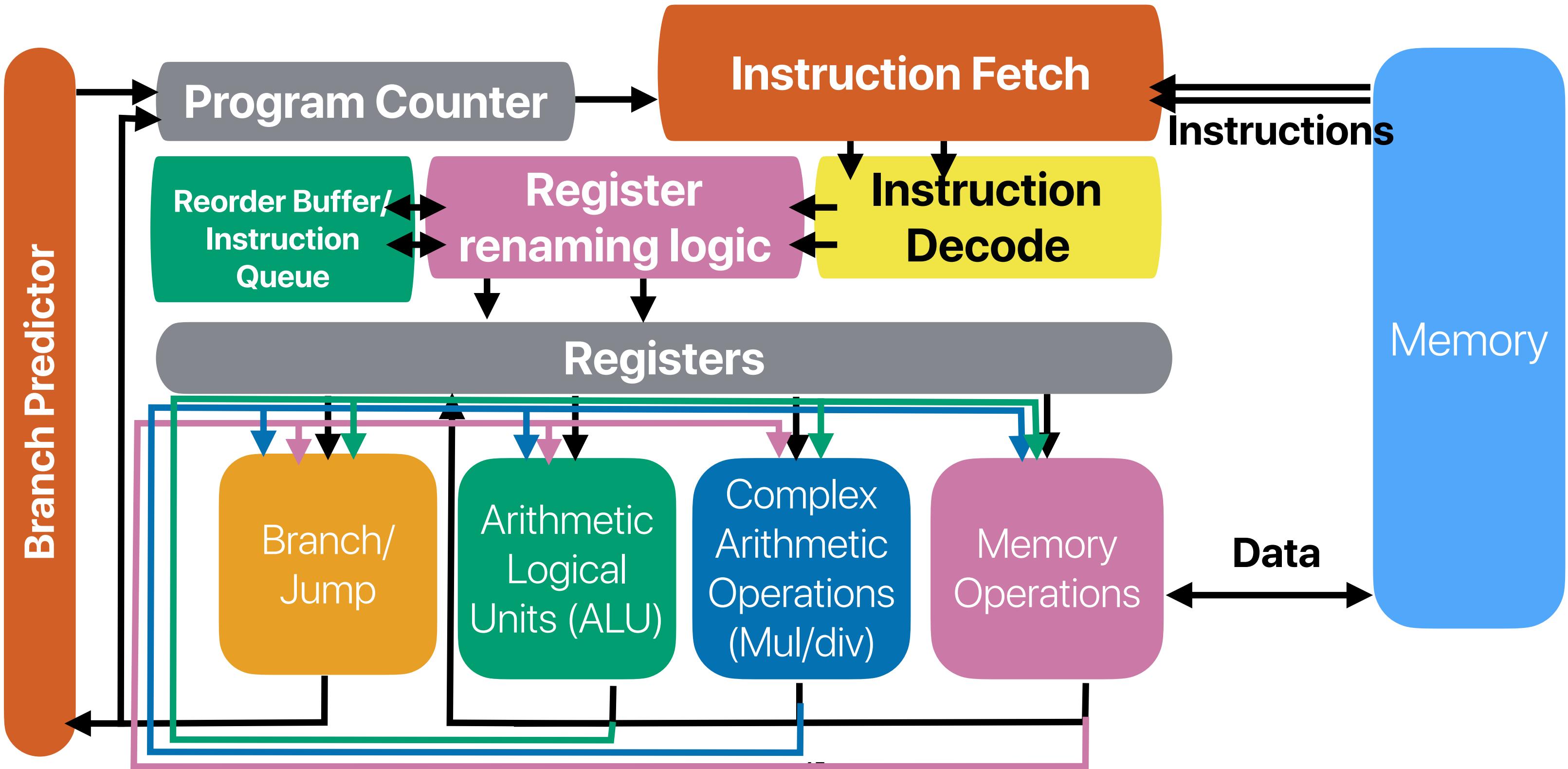
① movq 8(%rdi), %rdi
② addl \$1, %eax
③ testq %rdi, %rdi
④ jne .L3
⑤ movq 8(%rdi), %rdi
⑥ addl \$1, %eax
⑦ testq %rdi, %rdi
⑧ jne .L3
⑨ movq 8(%rdi), %rdi
⑩ addl \$1, %eax
⑪ testq %rdi, %rdi
⑫ jne .L3
⑬ movq 8(%rdi), %rdi
⑭ addl \$1, %eax
⑮ testq %rdi, %rdi
⑯ jne .L3
⑰ movl (%rdi), %ecx

By scheduling another running program instance (thread), the processor has 0 wasted issue slots!



- ① movl (%rdi), %ecx
- ② addq \$4, %rdi
- ③ addl %ecx, %eax
- ④ cmpq %rdx, %rdi
- ⑤ jne .L3
- ⑥ movl (%rdi), %ecx
- ⑦ addq \$4, %rdi
- ⑧ addl %ecx, %eax
- ⑨ cmpq %rdx, %rdi
- ⑩ jne .L3
- ⑪ movl (%rdi), %ecx
- ⑫ addq \$4, %rdi
- ⑬ addl %ecx, %eax
- ⑭ cmpq %rdx, %rdi
- ⑮ jne .L3
- ⑯ movl (%rdi), %ecx

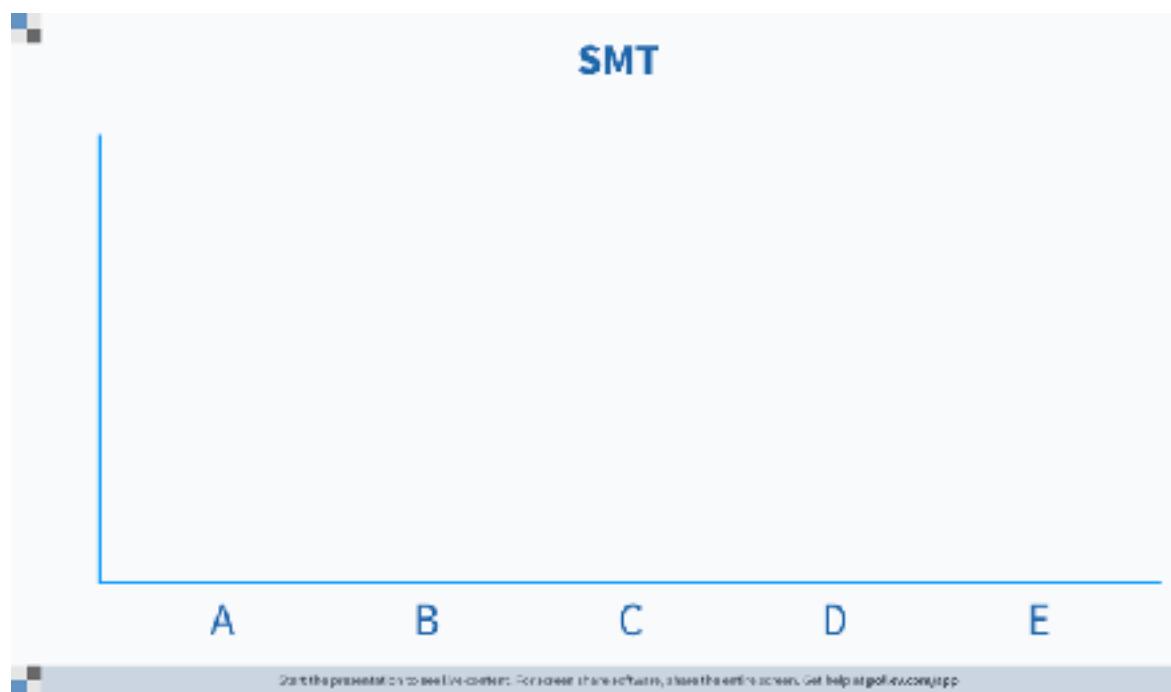
Register renaming



Architectural support for simultaneous multithreading

- To create an illusion of a multi-core processor and allow the core to run instructions from multiple threads concurrently, how many of the following units in the processor must be duplicated/extended?
 - ① Program counter
 - ② Register mapping tables
 - ③ Physical registers
 - ④ ALUs
 - ⑤ Data cache
 - ⑥ Reorder buffer/Instruction Queue

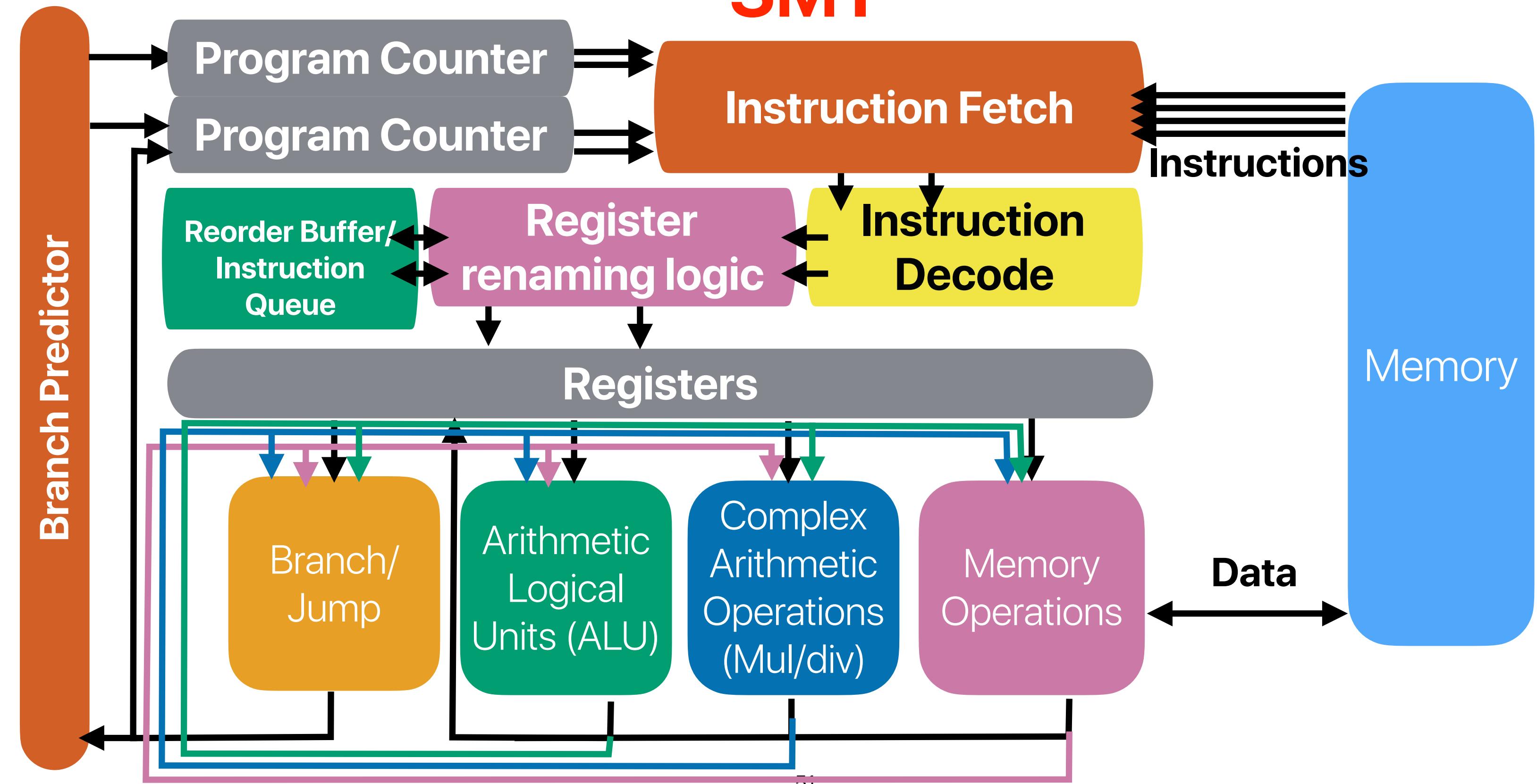
A. 2
B. 3
C. 4
D. 5
E. 6



Architectural support for simultaneous multithreading

- To create an illusion of a multi-core processor and allow the core to run instructions from multiple threads concurrently, how many of the following units in the processor must be duplicated/extended?
 - ① Program counter — **you need to have one for each context**
 - ② Register mapping tables — **you need to have one for each context**
 - ③ Physical registers — **you can share**
 - ④ ALUs — **you can share**
 - ⑤ Data cache — **you can share**
 - ⑥ Reorder buffer/Instruction Queue
A. 2 — **you need to indicate which context the instruction is from**
 - B. 3
 - C. 4
 - D. 5
 - E. 6

SMT



SMT

- Improve the throughput of execution
 - May increase the latency of a single thread
- Less branch penalty per thread
- Increase hardware utilization
- Simple hardware design: Only need to duplicate PC/Register Files
- Real Case:
 - Intel HyperThreading (supports up to two threads per core)
 - Intel Pentium 4, Intel Atom, Intel Core i7
 - AMD Ryzen (Zen microarchitecture)
 - If you see a processor with “threads” more than “cores”, that must be because of SMT!

Architecture:	x86_64
CPU op-mode(s):	32-bit, 64-bit
Byte Order:	Little Endian
Address sizes:	48 bits physical, 48 bits virtual
CPU(s):	16
On-line CPU(s) list:	0-15
Thread(s) per core:	2
Core(s) per socket:	8
Socket(s):	1
NUMA node(s):	1
Vendor ID:	AuthenticAMD
CPU family:	25
Model:	80
Model name:	AMD Ryzen 7 5700G with Radeon Graphics
Stepping:	0

Announcements

- Last Reading Quiz due next Tuesday
- Assignment #4 due 6/8
- If you submit iEVAL and submit the screenshot through gradescope, it counts as a “full-credit” notebook assignment

Computer Science & Engineering

203

つづく

