/INFOMOV/

Optimization & Vectorization

J. Bikker - Sep-Nov 2018 - Lecture 3: “Caching (1)”

Welcome!

Oops

“Doubles are slower than floats (4x)”

This statement is **mostly false**. The real story, CPU (win32, x64):

- A `float` takes 32-bit in memory, but gets promoted to 80 bits in an FPU register.
- A `double` takes 64-bit in memory, but gets promoted to 80 bits in an FPU register.
- A `long double` takes 64-bit in memory, but gets promoted to 80 bits in an FPU register.

Calculation time on 80-bit FPU registers does not depend on the source of the data.

The real story, GPU (Nvidia, AMD): https://www.geeks3d.com/20140305/amd-radeon-and-nvidia-geforce-fp32-fp64-gflops-table-computing

- Titan V: FP64 = 1/2 * FP32 (6900 vs 13800 GFLOPS)
- Titan X Pascal: FP64 = 1/32 * FP32 (350 vs 11300 GFLOPS) (same for all 10xx)
- Radeon RX Vega 64: FP64 = 1/16 * FP32 (790 vs 12700 GFLOPS)
- Radeon HD 7990: FP64 = 1/4 * FP32 (1946 vs 7782)

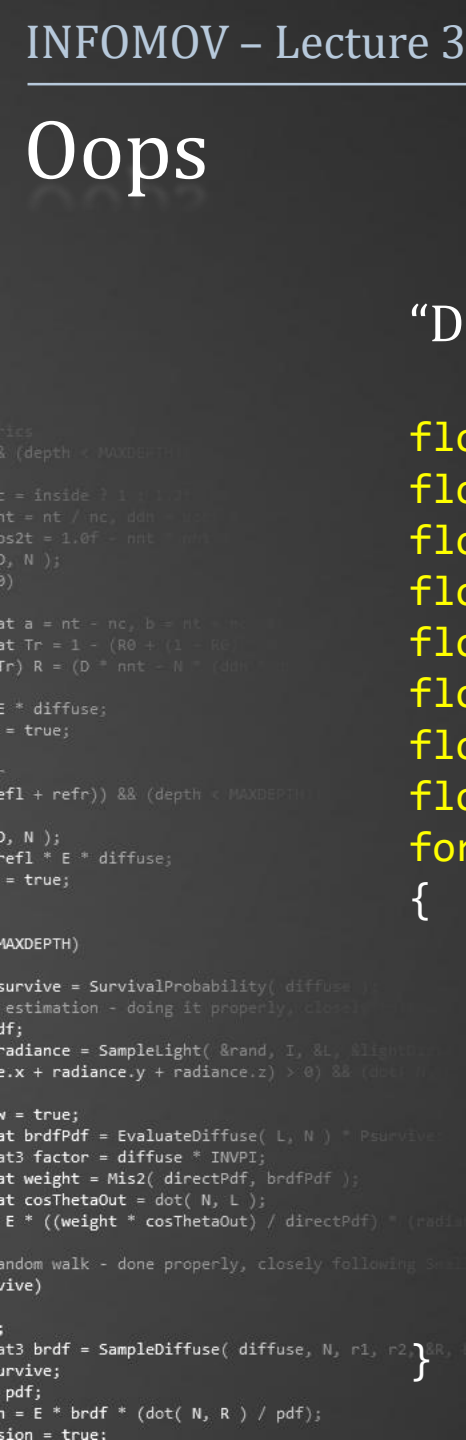
FP16 (GPU only): https://www.anandtech.com/show/10325/the-nvidia-geforce-gtx-1080-and-1070-founders-edition-review/5

- GTX 1080Ti: FP16 = 1/64 * FP32 (ouch)
- Radeon RX Vega 64: FP16 = 2 * FP32 (!)

Oops

“Doubles are slower than floats (4x)”

```
float v0 = 1;                          fld1
float v1 = 1;                          fld st(0)
float v2 = 1;                          fld st(1)
float v3 = 1;                          fld st(2)
float v4 = 1;                          fld st(3)
float v5 = 1;                          fld st(4)
float v6 = 1;                          fld st(5)
float v7 = 1;                          fld st(6)
for (int i = 0; i < 2000000; i++)
{
   v0 *= 1.00001f;                     fmul st(7),st ; fxch st(7) ; fstp [v0]
   v1 *= 1.00001f;                     fxch st(5) ; fmul st,st(6)
   v2 *= 1.00001f;                     fxch st(4) ; fmul st,st(6)
   v3 *= 1.00001f;                     fxch st(3) ; fmul st,st(6)
   v4 *= 1.00001f;                     fxch st(2) ; fmul st,st(6)
   v5 *= 1.00001f;                     fxch st(1) ; fmul st,st(6)
   v6 *= 1.00001f;                     fxch st(5) ; fmul st,st(6)
   v7 *= 1.00001f;                     fld [v7] fmul st,st(7) fstp [v7]
}
```



Today’s Agenda:

- The Problem with Memory
- Cache Architectures
- Practical Assignment 2

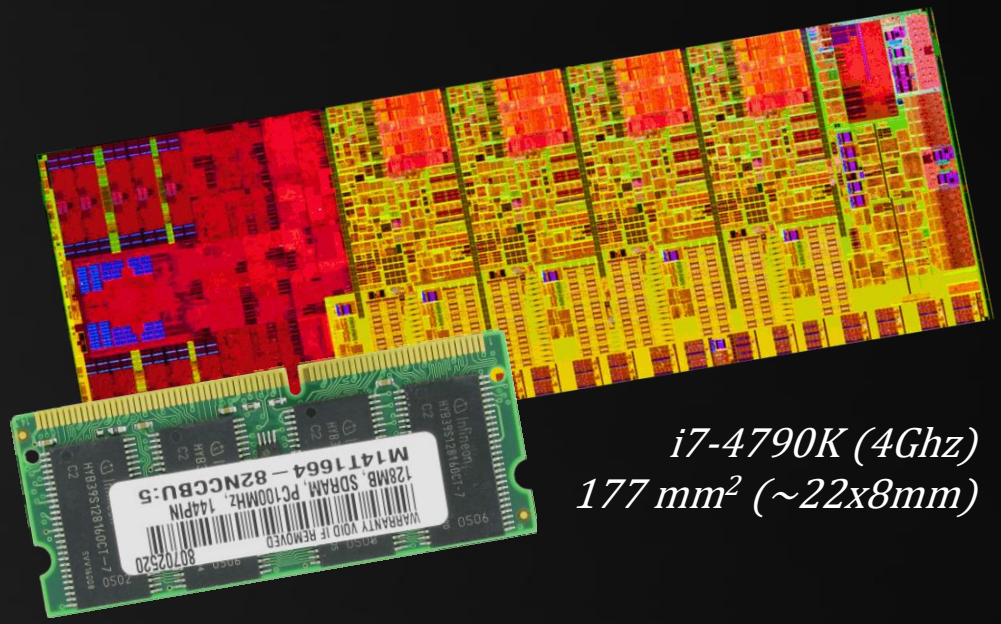
Introduction

Feeding the Beast

Let’s assume our CPU runs at 4Ghz.
What is the maximum physical distance between memory and CPU if we want to retrieve data every cycle?

Speed of light (vacuum): 299,792,458 m/s
Per cycle: ~0.075 m
➔ ~**3.75cm** back and forth.

In other words: we cannot physically query RAM fast enough to keep a CPU running at full speed.



i7-4790K (4Ghz)
177 mm² (~22x8mm)

Introduction

Feeding the Beast

Sadly, we can’t just divide by the physical distance between CPU and RAM to get the cycles required to query memory.

Factors include (stats for DDR4-3200/PC4-25600):

- RAM runs at a much lower clock speed than the CPU
  - 25600 here means: theoretical bandwidth in MB/s
  - 3200 is the number of transfers per second (1 transfer=64bit)
  - We get two transfers per cycle, so actual I/O clock speed is 1600Mhz
  - DRAM cell array clock is ~1/4th of that: 400Mhz.
- Latency between query and response: 20-24 cycles.

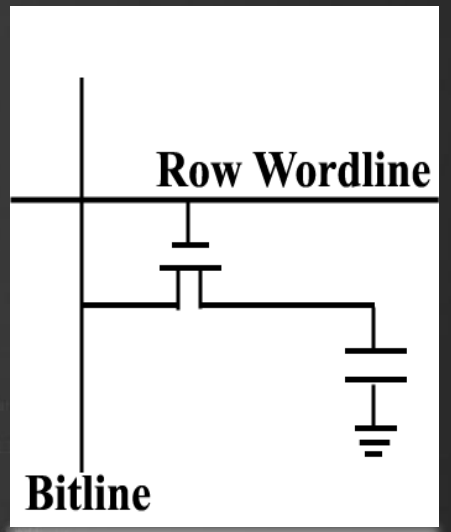
Introduction

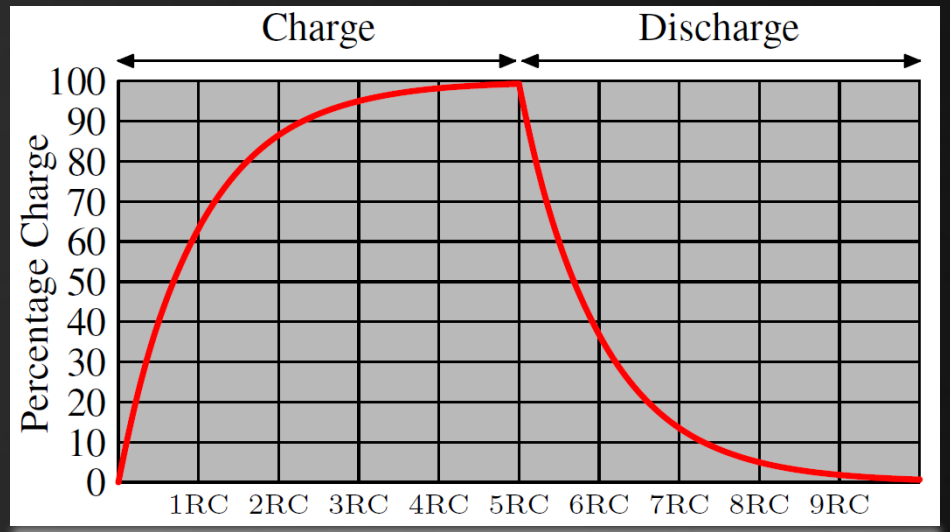
Feeding the Beast

Sadly, we can’t just divide by the physical distance between CPU and RAM to get the cycles required to query memory.

Factors include (stats for DDR4-3200/PC4-25600):

- Latency between query and response: 20-24 cycles.





Introduction

Feeding the Beast

Sadly, we can’t just divide by the physical distance between CPU and RAM to get the cycles required to query memory.

Additional delays may occur when:

- Other devices than the CPU access RAM;
- DRAM must be refreshed every 64ms due to leakage.

For a processor running at 2.66GHz, latency is roughly 110-140 CPU cycles.

Details in: “What Every Programmer Should Know About Memory”, chapter 2.

Introduction

Feeding the Beast

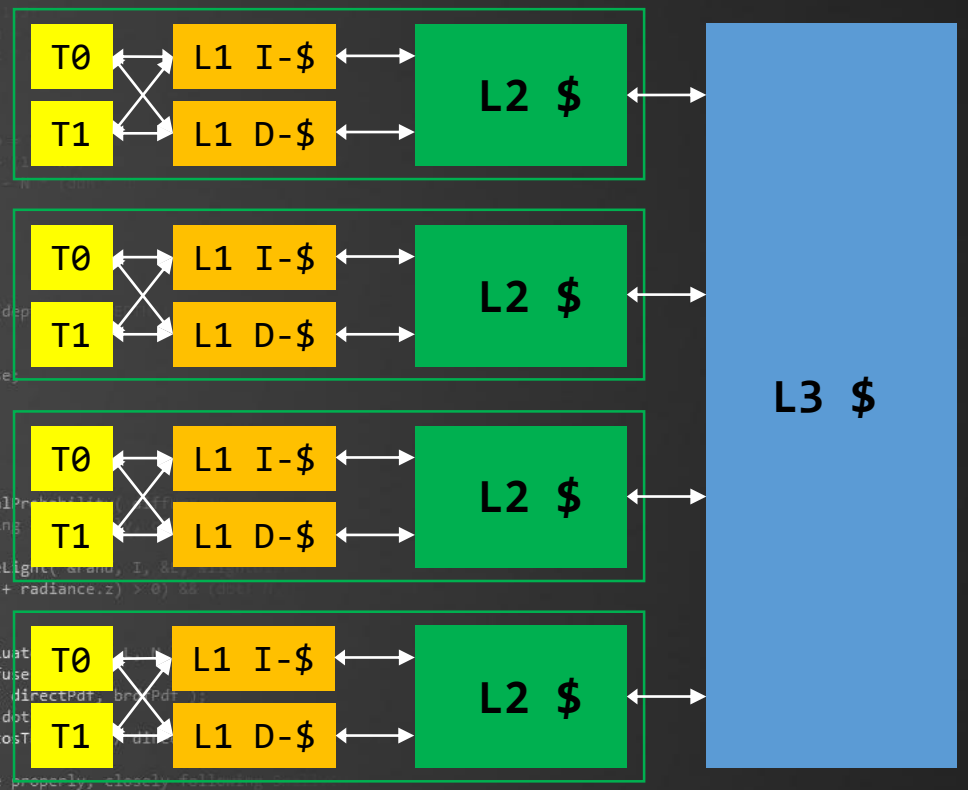
“We cannot physically query RAM fast enough to keep a CPU running at full speed.”

How do we overcome this?

We keep a copy of frequently used data in fast memory, close to the CPU: the ***cache***.

Introduction

The Memory Hierarchy – Core i7-9xx (4 cores)

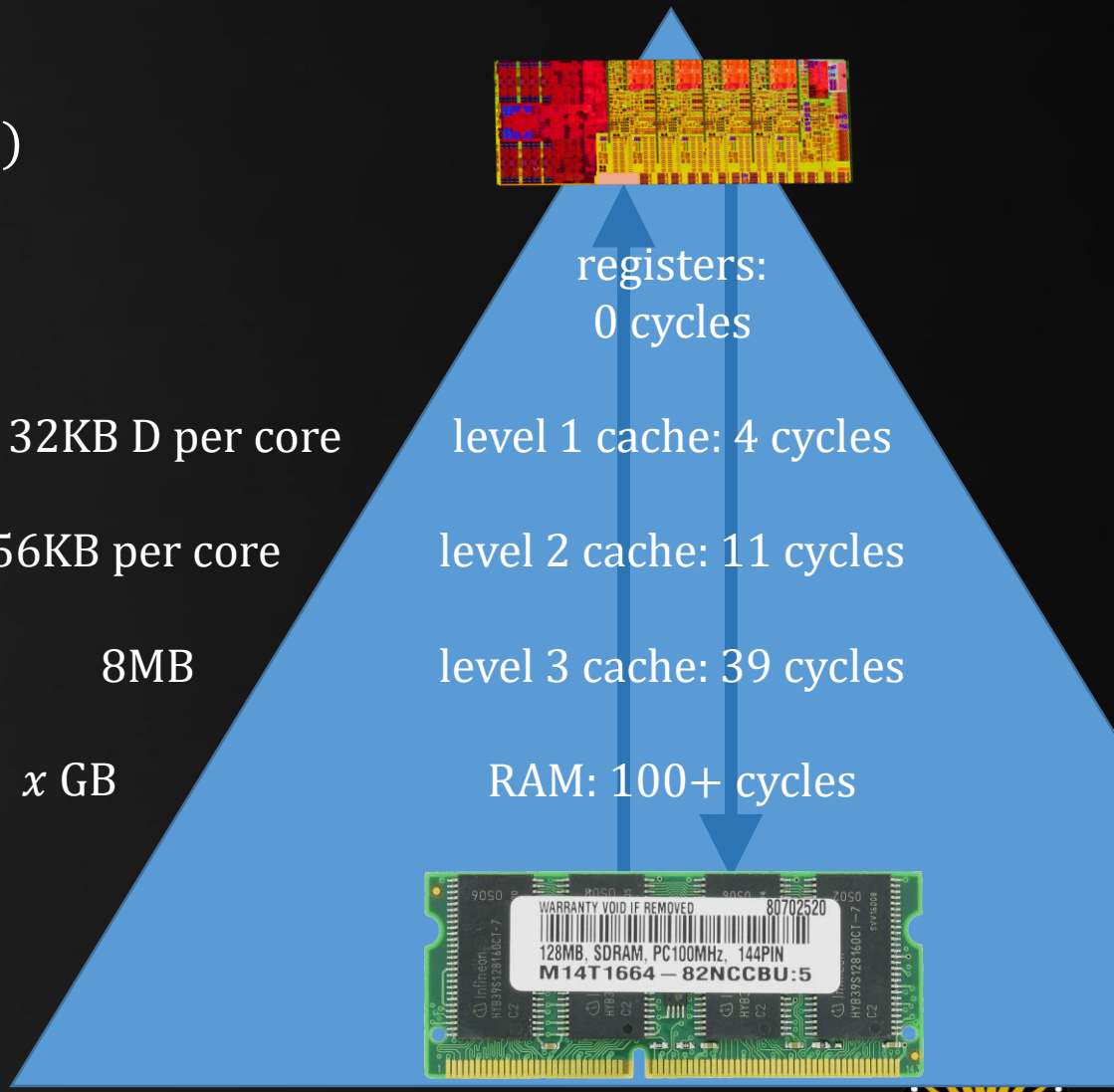


32KB I / 32KB D per core

256KB per core

8MB

x GB



Introduction

Caches and Optimization

Considering the cost of RAM vs L1 cache access, it is clear that the cache is an important factor in code optimization:

- Fast code communicates mostly with the caches
- We still need to get data into the caches
- But ideally, only once.

Therefore:

- The working set must be small;
- Or we must maximize *data locality*.

Today’s Agenda:

- The Problem with Memory
- Cache Architectures
- Practical Assignment 2

Architectures

Cache Architecture

The simplest caching scheme is the *fully associative cache*.

```
struct CacheLine
{
    uint address; // 32-bit for 4G
    uchar data;
    bool valid;
};
CacheLine cache[256];
```

This cache holds 256 bytes.

address	data	valid
0x00000000	0xFF	0
0x00000000	0xFF	0
0x00000000	0xFF	0
0x00000000	0xFF	0
0x00000000	0xFF	0
…	…	…
0x00000000	0xFF	0

Notes on this layout:

- We will rarely need 1 byte at a time
- So, we switch to 32bit values
- We will rarely read those at odd addresses
- So, we drop 2 bits from the address field.

Architectures

Cache Architecture

The simplest caching scheme is the *fully associative cache*.

```
struct CacheLine
{
    uint tag;      // 30 bit for 4G
    uint data;
    bool valid, dirty;
};
CacheLine cache[64];
```

This cache holds 64 dwords (256 bytes).

tag	data	valid	dirty
0x00000000	0xFFFFFFFF	0	0
0x00000000	0xFFFFFFFF	0	0
0x00000000	0xFFFFFFFF	0	0
0x00000000	0xFFFFFFFF	0	0
0x00000000	0xFFFFFFFF	0	0
…	…		
0x00000000	0xFFFFFFFF	0	0

Architectures

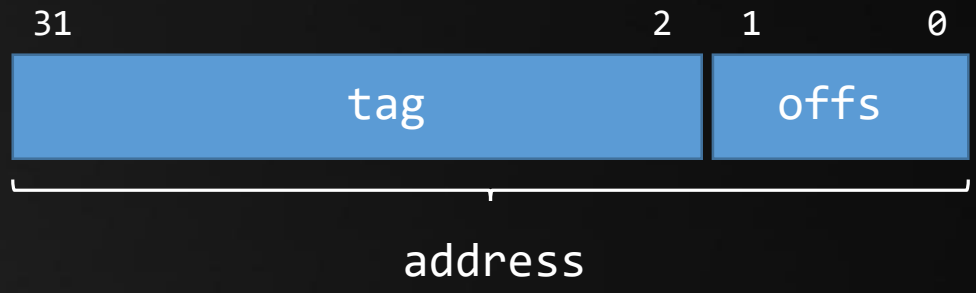
Cache Architecture

The simplest caching scheme is the *fully associative cache*.

```
struct CacheLine
{
   uint tag;      // 30 bit for 4G
   uint data;
   bool valid, dirty;
};
CacheLine cache[64];
```

This cache holds 64 dwords (256 bytes).

Single-byte read operation:




```
for ( int i = 0; i < 64; i++ )
   if (cache[i].valid)
      if (cache[i].tag == tag)
         return cache[i].data[offs];

uint d = RAM[tag].data; // cache miss

WriteToCache( tag, d );

return d[offs];
```

Architectures

Cache Architecture

The simplest caching scheme is the *fully associative cache*.

```
struct CacheLine
{
   uint tag;      // 30 bit for 4G
   uint data;
   bool valid, dirty;
};
CacheLine cache[64];
```

This cache holds 64 dwords (256 bytes).

Single-byte write operation:

```
for ( int i = 0; i < 64; i++ )
   if (cache[i].valid)
      if (cache[i].tag == a)
         cache[i].data[offs] = d;
         cache[i].dirty = true;
         return;

for ( int i = 0; i < 64; i++ )
   if (!cache[i].valid)
      cache[i].tag = a;
      cache[i].data[offs] = d;
      cache[i].valid|dirty = true;
      return;

i = BestSlotToOverwrite();

if (cache[i].dirty) SaveToRam(i);
cache[i].tag = a;
cache[i].data[offs] = d;
cache[i].valid|dirty = true;
```

One problem remains… We store one byte, but the slot stores 4. What should we do with the other 3?

Architectures

BestSlotToOverwrite() ?

The best slot to overwrite is the one that will not be needed for the longest amount of time. This is known as Bélády’s algorithm, or the *clairvoyant* algorithm.

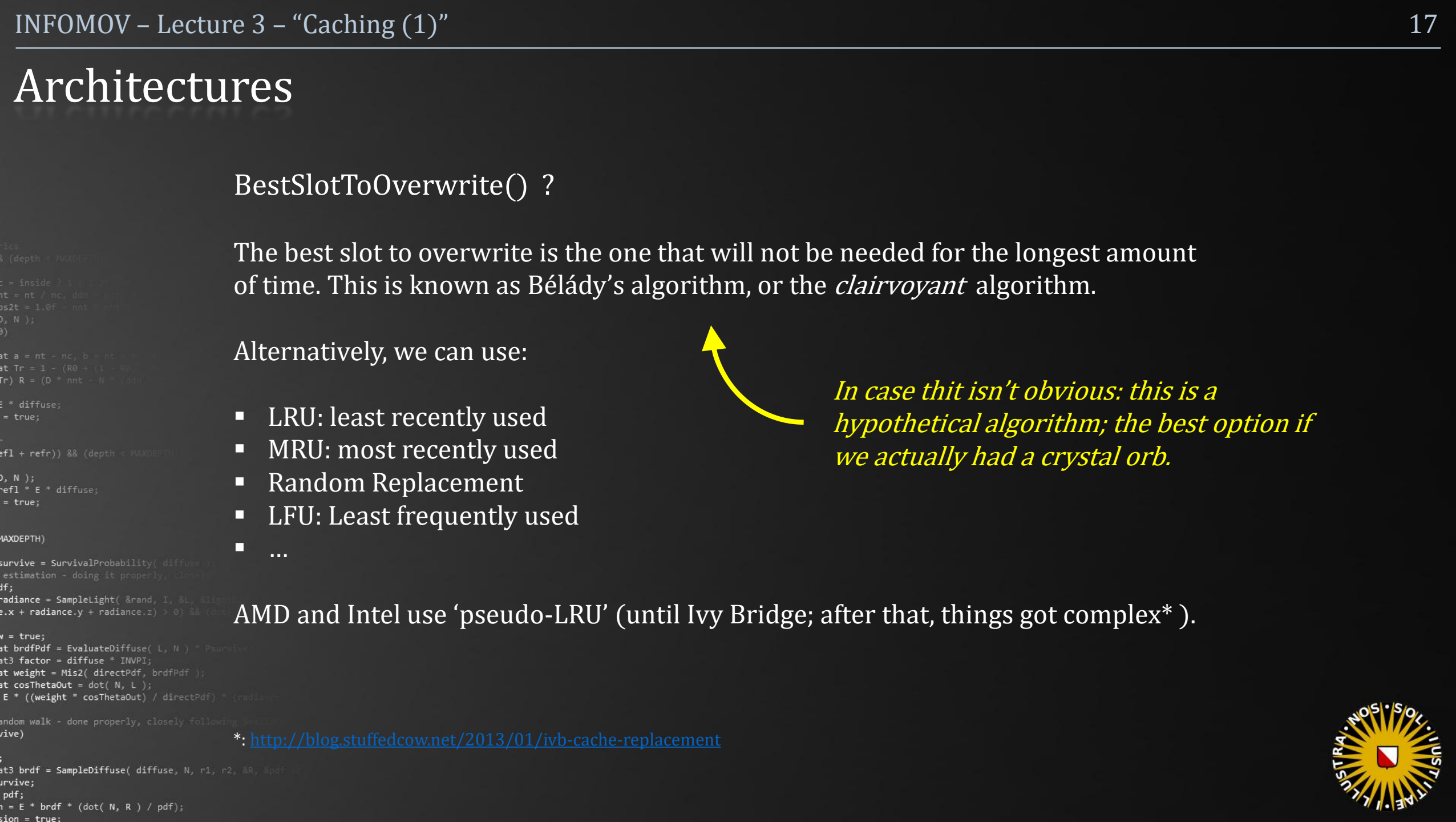
Alternatively, we can use:

- LRU: least recently used
- MRU: most recently used
- Random Replacement
- LFU: Least frequently used
- …

In case thit isn’t obvious: this is a hypothetical algorithm; the best option if we actually had a crystal orb.

AMD and Intel use ‘pseudo-LRU’ (until Ivy Bridge; after that, things got complex*).

*: http://blog.stuffedcow.net/2013/01/ivb-cache-replacement



Architectures

The Problem with Being Fully Associative

Read / Write using a fully associative cache is O(N): we need to scan each entry. This is not practical for anything beyond 16~32 entries.

An alternative scheme is the *direct mapped cache*.

Architectures

Direct Mapped Cache

```
struct CacheLine
{
   uint tag;      // 24 bit for 4G
   uint data;
   bool dirty, valid;
};
CacheLine cache[64];
```

This cache again holds 256 bytes.

In a direct mapped cache, each address can only be stored in a single cache line.

Read/write access is therefore O(1).

For a cache consisting of 64 cache lines:

- Bit 0 and 1 still determine the offset within a slot;
- 6 bits are used to determine which slot to use;
- The remaining 24 bits form the tag.

Architectures

Direct Mapped Cache

In general:

$N = \log_2(cache\ line\ width)$
$M = \log_2(number\ of\ slots\ in\ the\ cache)$

- Bits 0..N-1 are used as offset in a cache line;
- Bits N..M-1 are used as slot index;
- Bits M..31 are used as tag.

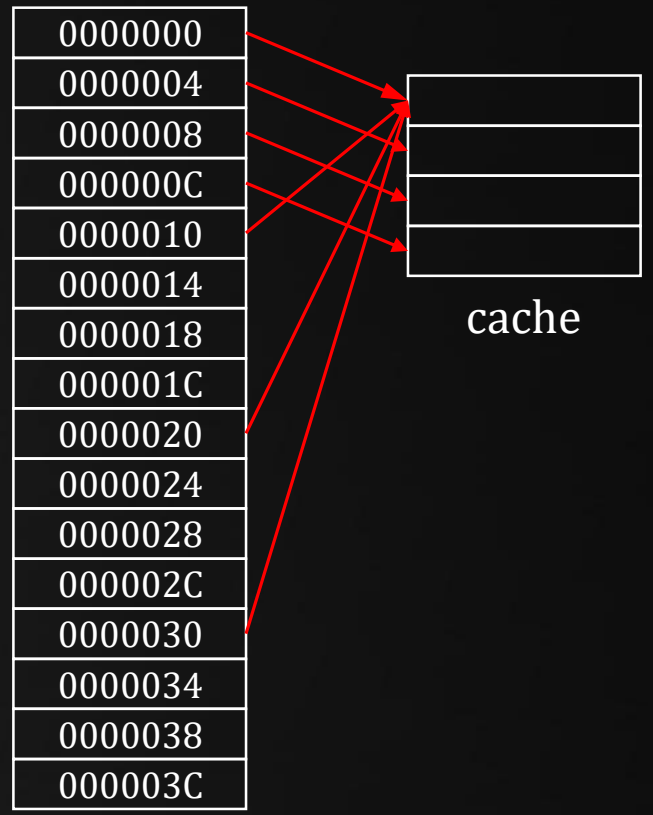
Architectures

The Problem with Direct Mapping

In this type of cache, each address maps to a single cache line, leading to O(1) access time. On the other hand, a single cache line ‘represents’ multiple memory addresses.

This leads to a number of issues:

- A program may use two variables that occupy the same cache line, resulting in frequent cache misses (collisions);
- A program may heavily use one part of the cache, and underutilize another.



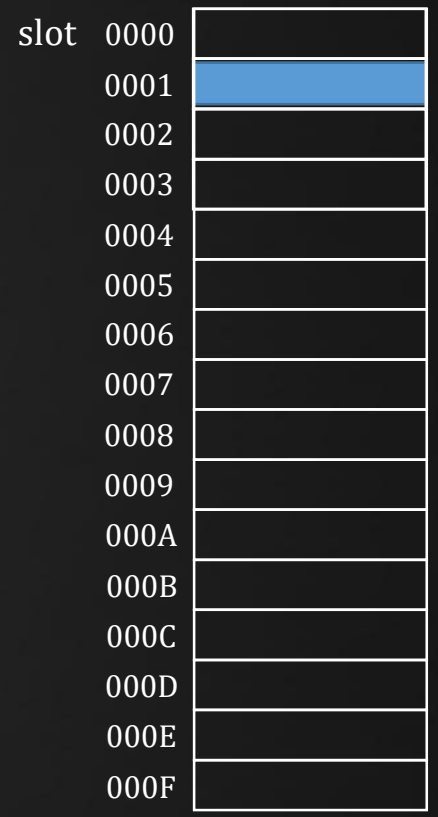
Architectures

N-Way Set Associative Cache

```
struct CacheLine
{
   uint tag;
   uint data;
   bool valid, dirty;
};
CacheLine cache[16][4];
```

This cache again holds 256 bytes.

In an N-way set associative cache, we use N slots (cache lines) per set.



Architectures

N-Way Set Associative Cache

```
struct CacheLine
{
   uint tag;  // 28 bit for 4G
   uint data;
   bool valid, dirty;
};
CacheLine cache[16][4];
```

This cache again holds 256 bytes.

In an N-way set associative cache, we use N slots (cache lines) per set.

When reading / writing data, we check each of the N slots that may contain the data.

Example: Address 0x00FF1004

Offset: lowest 2 bits ➔ 0.
Set: next 2 bits ➔ 1.
Tag: remaining bits.

Architectures

Caching Architectures

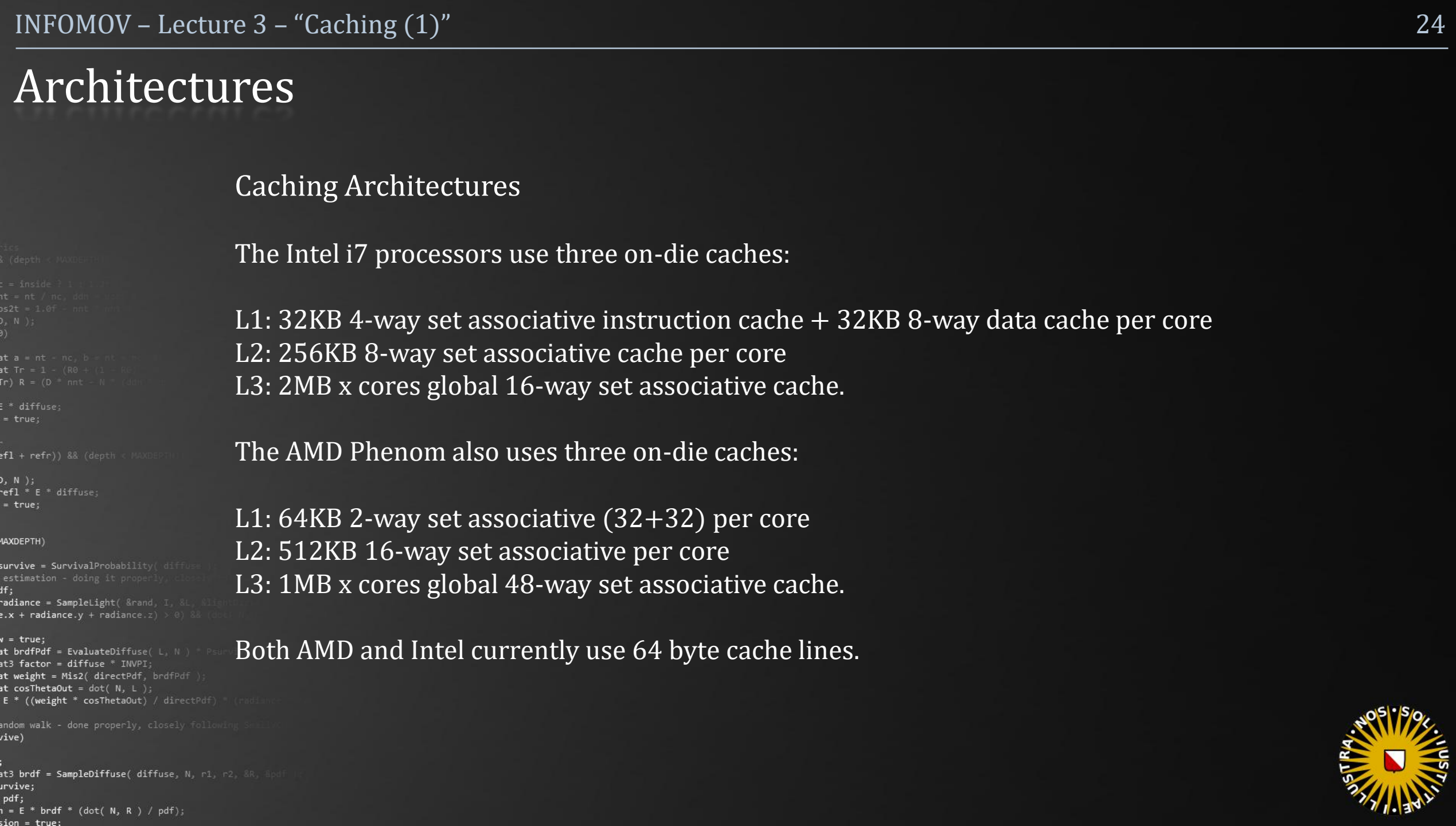
The Intel i7 processors use three on-die caches:

L1: 32KB 4-way set associative instruction cache + 32KB 8-way data cache per core
L2: 256KB 8-way set associative cache per core
L3: 2MB x cores global 16-way set associative cache.

The AMD Phenom also uses three on-die caches:

L1: 64KB 2-way set associative (32+32) per core
L2: 512KB 16-way set associative per core
L3: 1MB x cores global 48-way set associative cache.

Both AMD and Intel currently use 64 byte cache lines.

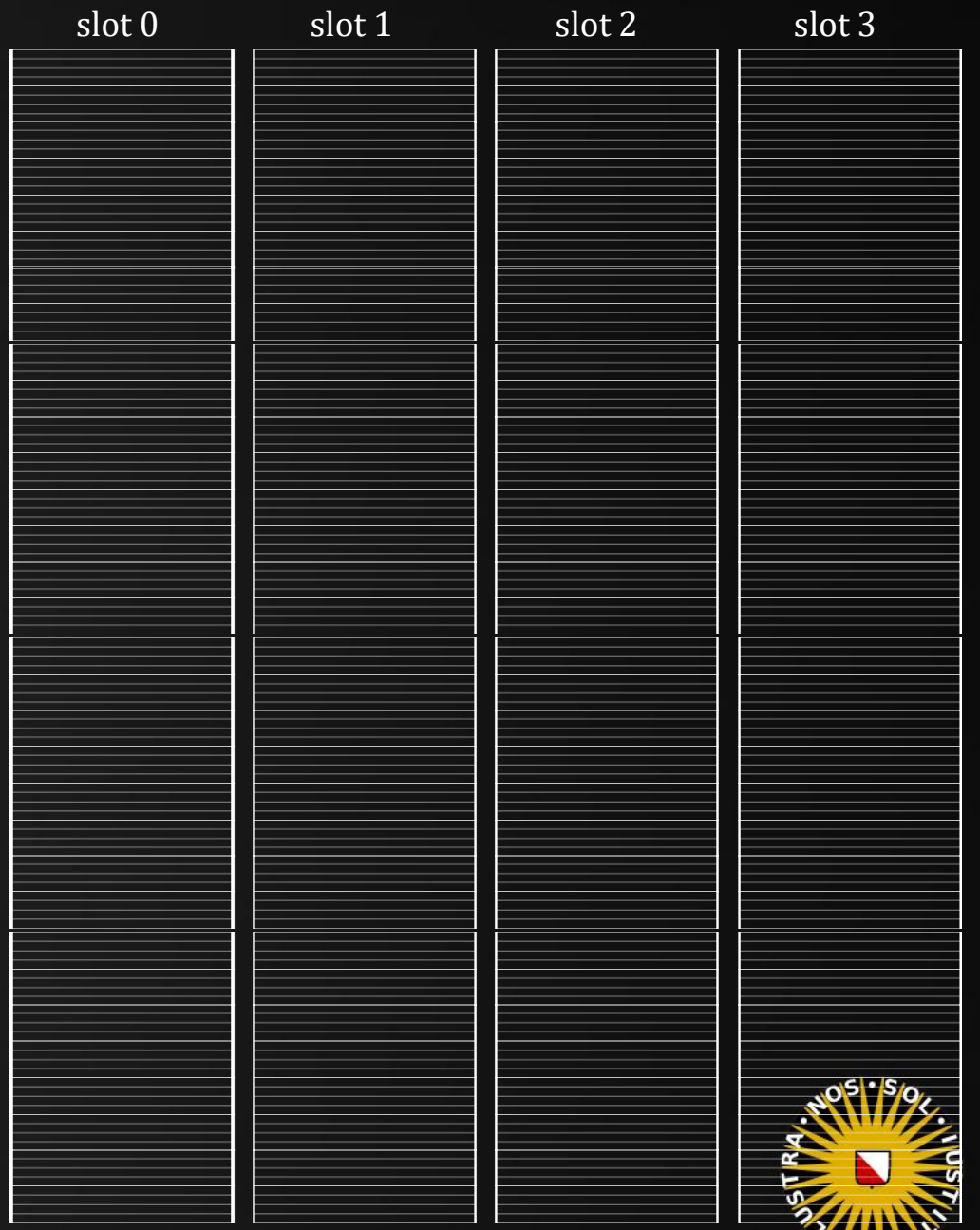


Architectures

32KB, 4-Way Set Associative Cache

```
struct CacheLine
{
   uint tag;  // 19 bit for 4G
   uchar data[64];
   bool valid, dirty;
};
CacheLine cache[128][4];
```

This cache holds 32768 bytes in 512 cachelines, organized in 128 sets of 4 cachelines.



Today’s Agenda:

- The Problem with Memory
- Cache Architectures
- Practical Assignment 2

Assignment 2

Second Practical Assignment: Create a Cache Simulator

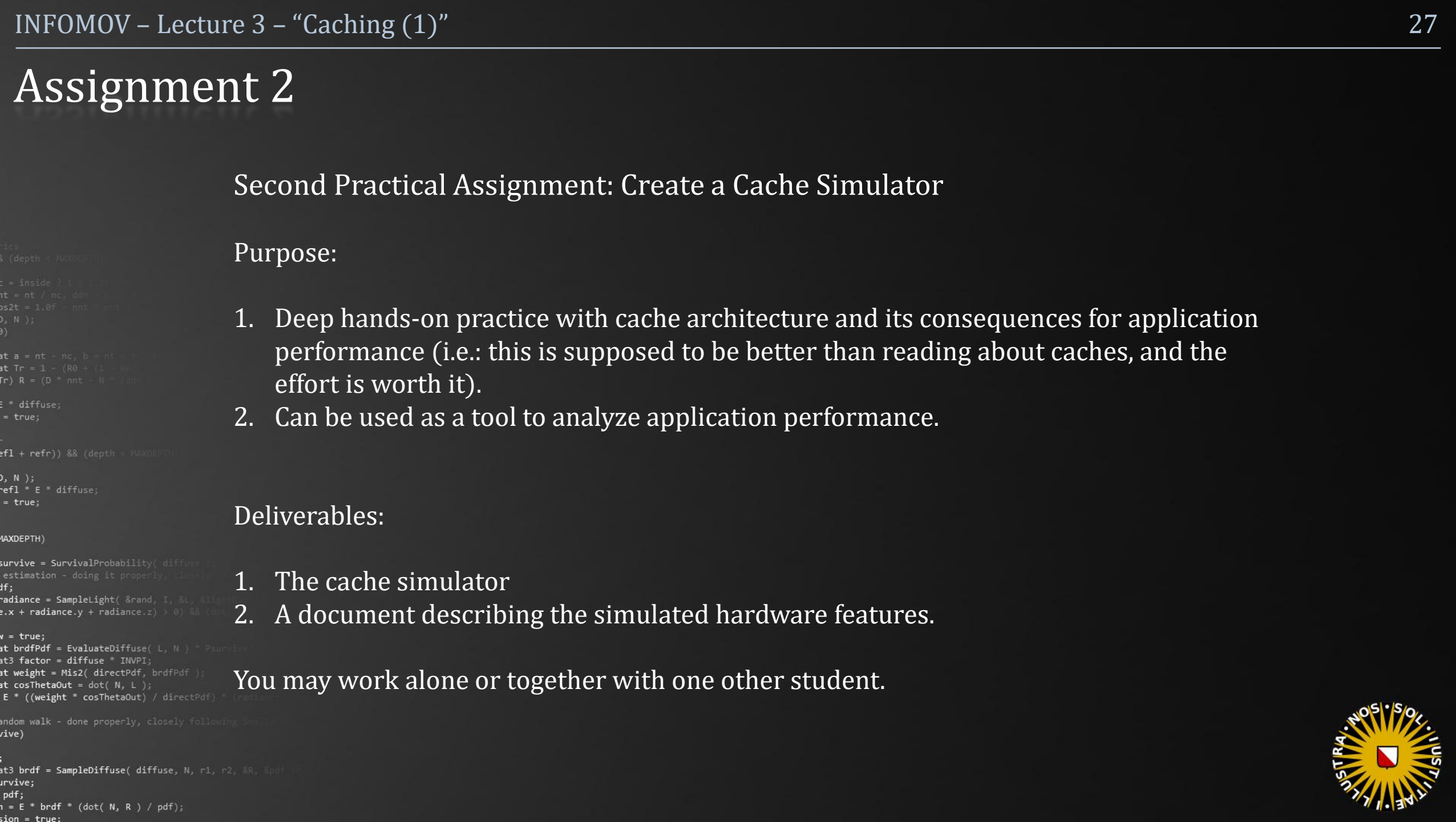
Purpose:

1. Deep hands-on practice with cache architecture and its consequences for application performance (i.e.: this is supposed to be better than reading about caches, and the effort is worth it).
2. Can be used as a tool to analyze application performance.

Deliverables:

1. The cache simulator
2. A document describing the simulated hardware features.

You may work alone or together with one other student.



Assignment 2

Second Practical Assignment: Create a Cache Simulator

Details, minimum requirements (for a 6):

1. implement a correct set associative cache within the supplied demo application;
2. implement a reasonable eviction policy. This requires some research.

Optional (towards a 9):

3. cover the full L1-L2-L3 cache hierarchy;
4. experiment with various eviction policies;
5. visualize cache efficiency over time.

Limitations:

- you may assume a single core (i.e. cache coherency and shared L3 doesn’t have to be simulated);
- the simulator doesn’t have to be efficient (memory / speed wise).

Assignment 2

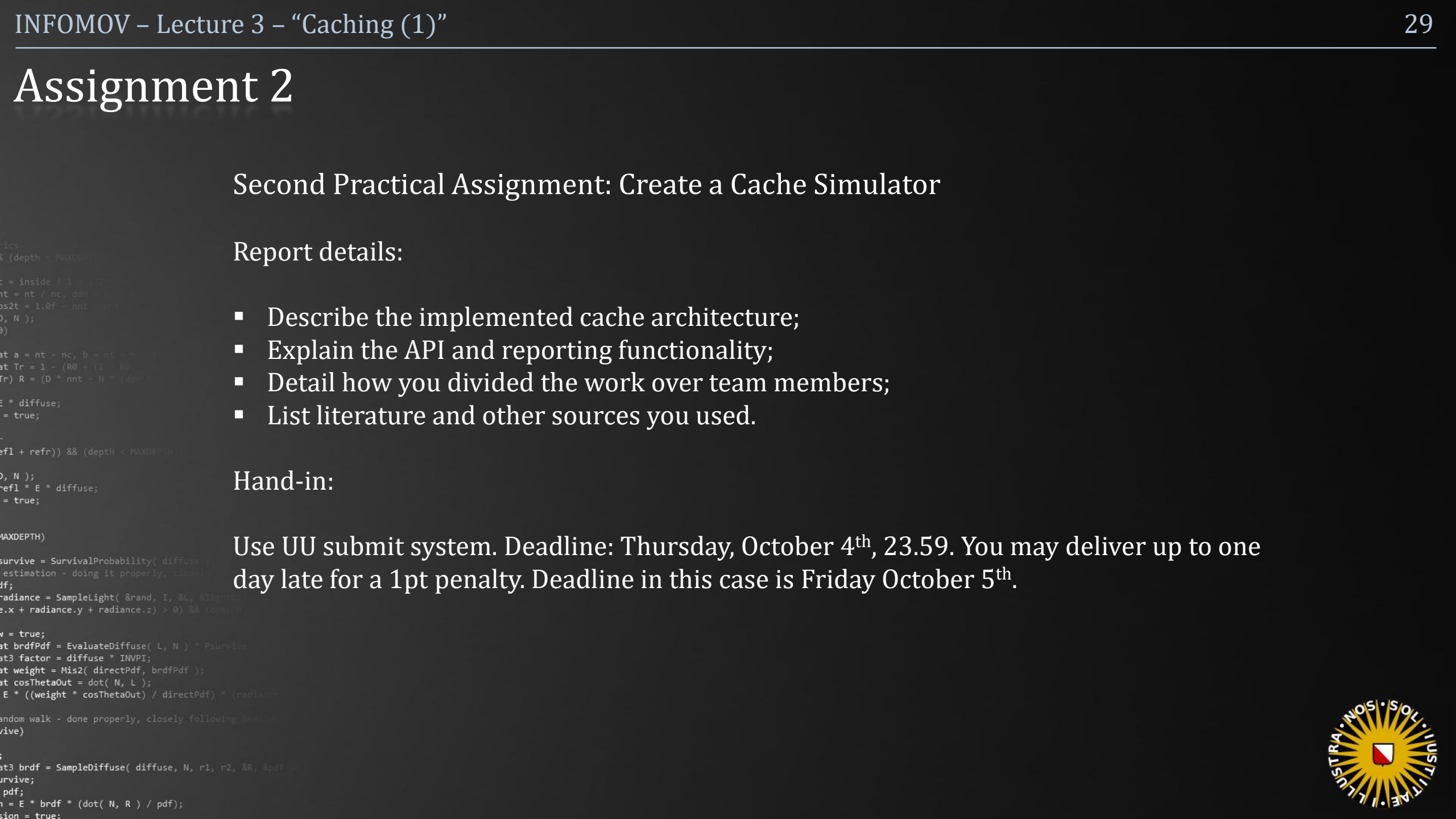
Second Practical Assignment: Create a Cache Simulator

Report details:

- Describe the implemented cache architecture;
- Explain the API and reporting functionality;
- Detail how you divided the work over team members;
- List literature and other sources you used.

Hand-in:

Use UU submit system. Deadline: Thursday, October 4th, 23.59. You may deliver up to one day late for a 1pt penalty. Deadline in this case is Friday October 5th.



Today’s Agenda:

- The Problem with Memory
- Cache Architectures
- Practical Assignment 2

/INFOMOV/

END of “Caching (1)”

next lecture: “Caching (2)”