Parallel Computing & OpenMP Introduction

Luca Tornatore - I.N.A.F.

"**Foundation of HPC" course**

DATA SCIENCE &
SCIENTIFIC COMPUTING
2019-2020 @ Università di Trieste

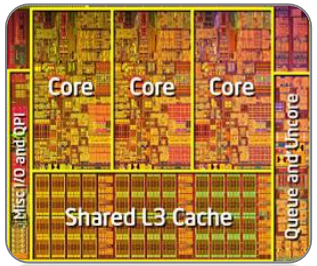
Outline

Introduction

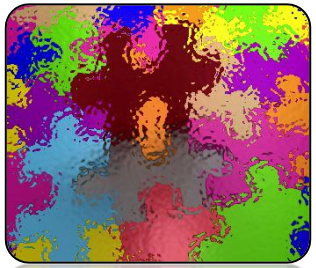
Parallel Computing

Intro to OpenMP

Introduction Outline



The race to multicore



Intro to Parallel Computing

Parallel Performance

Warm-up

A quick recap of what we have seen in the Optimization part …

"CRUCIAL PROBLEMS that we can only hope to address computationally REQUIRE US TO DELIVER EFFECTIVE COMPUTING POWER ORDERS-OF-MAGNITUDE GREATER THAN WE CAN DEPLOY TODAY."

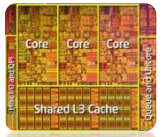
DOE's *Office of Science, 2012*

Why there is no more “free lunch”?

Applications no longer get more performance for free without significant re-design, since 15 years

Since 15 years, the gain in performance is essentially due to **fundamentally different factors**:

1. Multi-core + **Multi-threads**
2. Enlarging/improving **cache**
3. Hyperthreading *(smaller contribution)*



Race to Multicore

Why there is no more “free lunch”?

For instance:

2 Cores at 3GHz are basically 1 Core at 6GHz.. ?

False

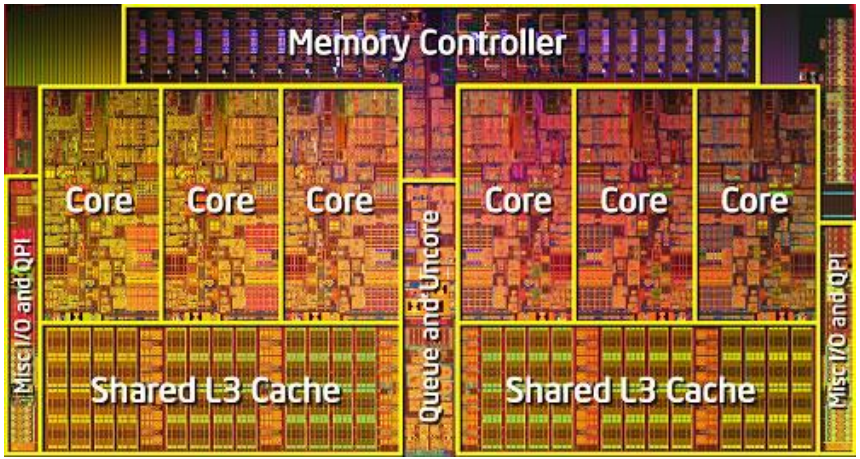
- × Cores coordination for cache-coherence
- × Threads coordination
- × Memory access
- × Increased algorithmic complexity

Since 15 years, the gain in performance is essentially due to **fundamentally different factors**:

1. Multi-core + **Multi-threads**
2. Enlarging/improving **cache**
3. Hyperthreading *(smaller contribution)*

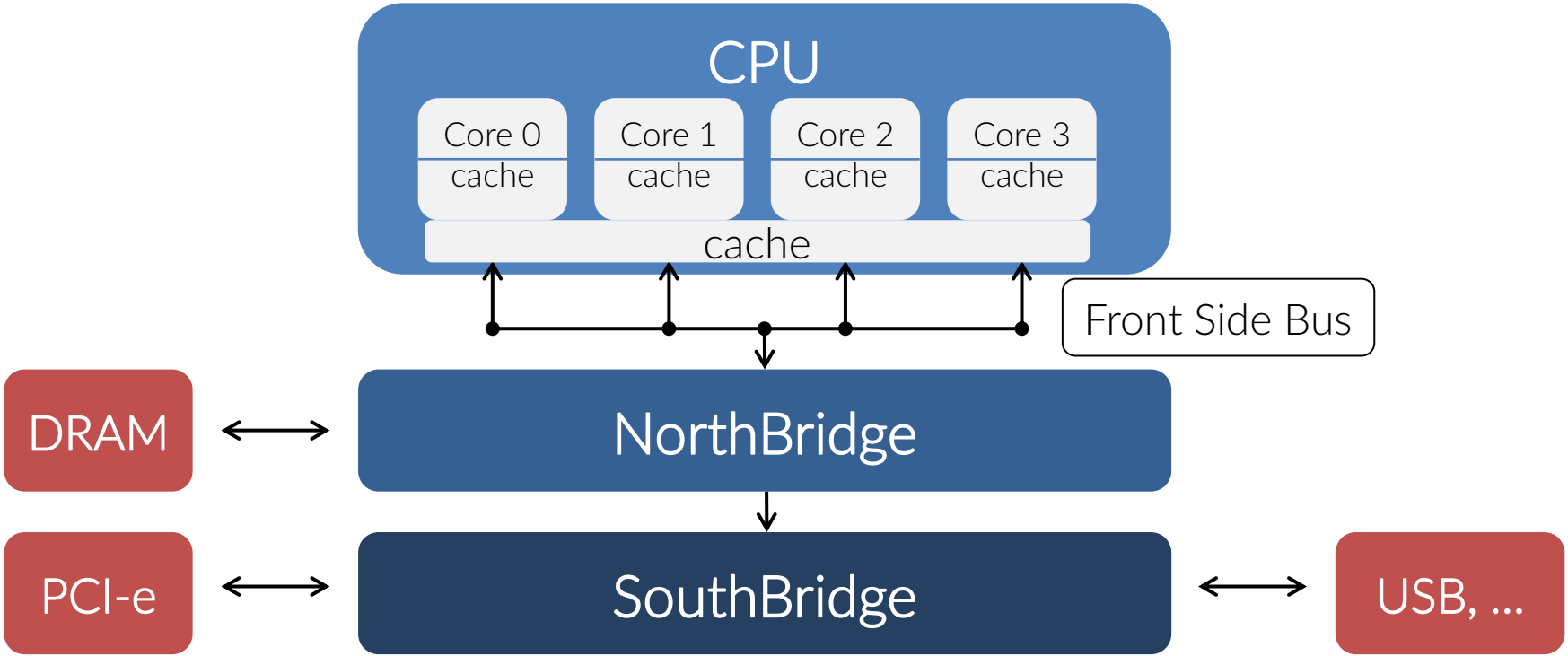
Back to the future

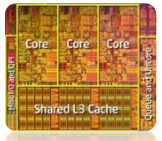
Message I
Many-cores CPUs are here to stay



- Concurrency-based model programming
  (different than both *parallel* and *ILP*): work subdivision in as many independent tasks as possible
- Specialized, heterogeneous cores
- Multiple memory hierarchies

The typical UMA architecture

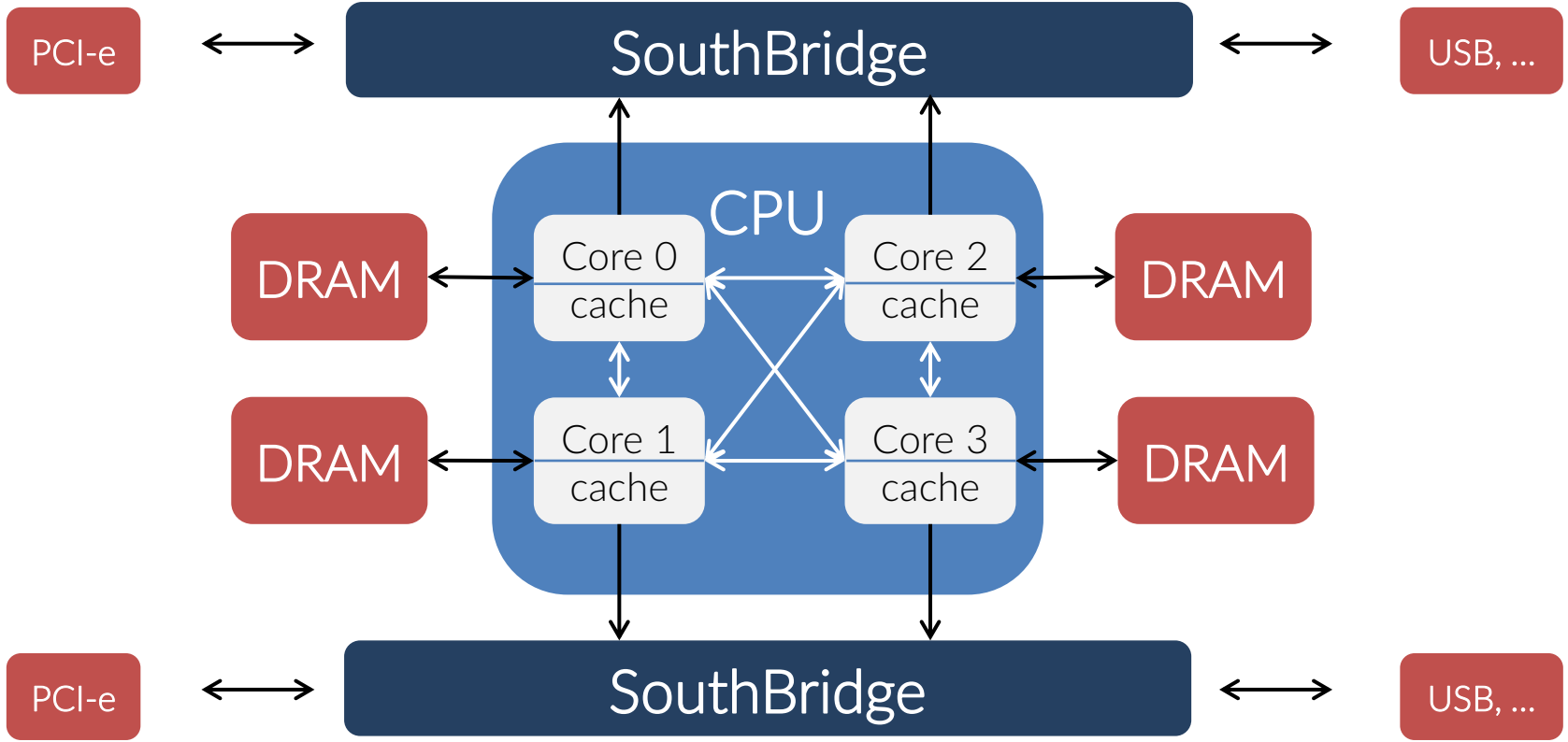




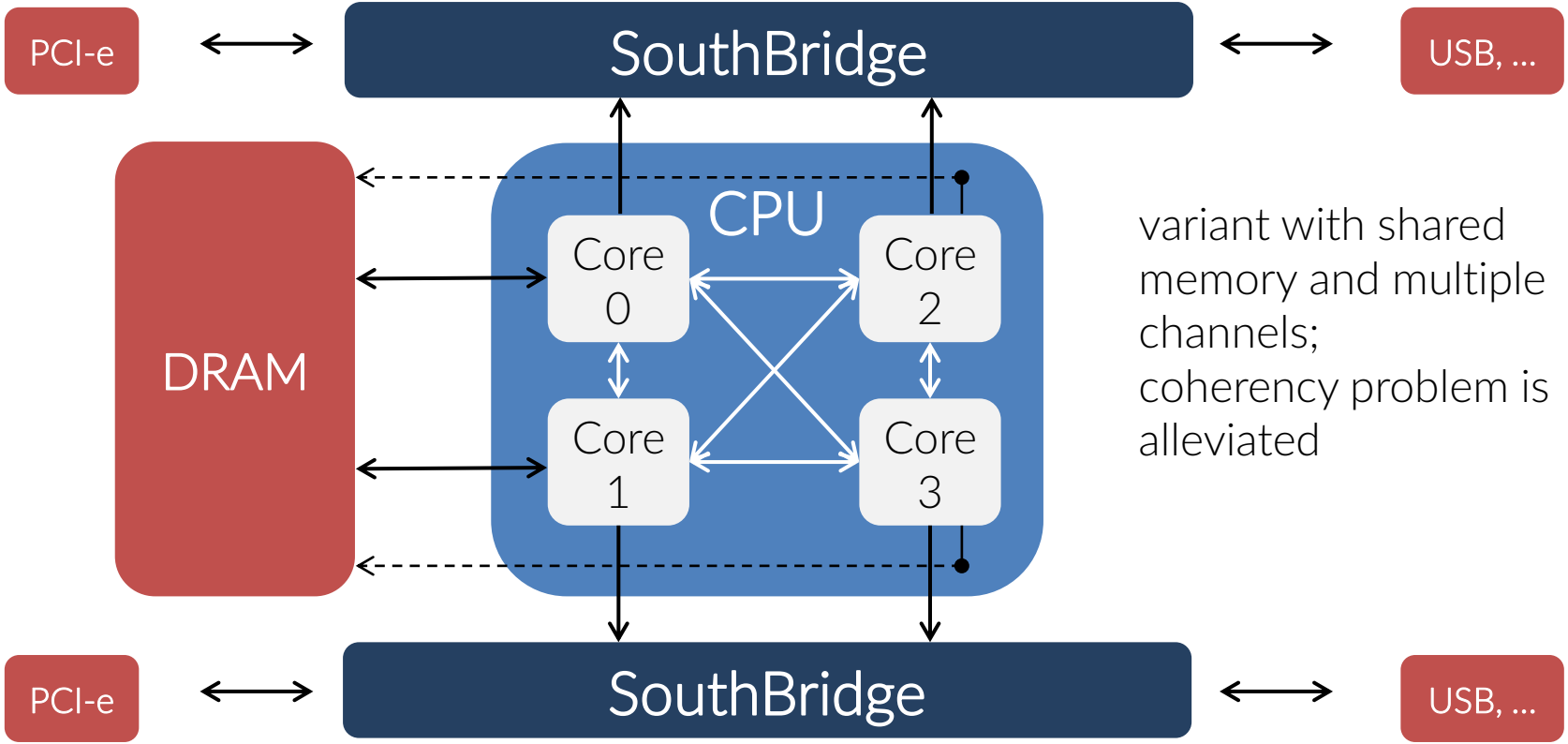
The typical UMA architecture

- The RAM can be accessed by one core at a time
  - this lowers the effective bandwidth
  - data coherence is easier
- The faster SRAM was introduced as caches to keep up with the increase of cores' clock
- FSB and RAM access is the main bottleneck

The typical NUMA architecture



The typical NUMA architecture



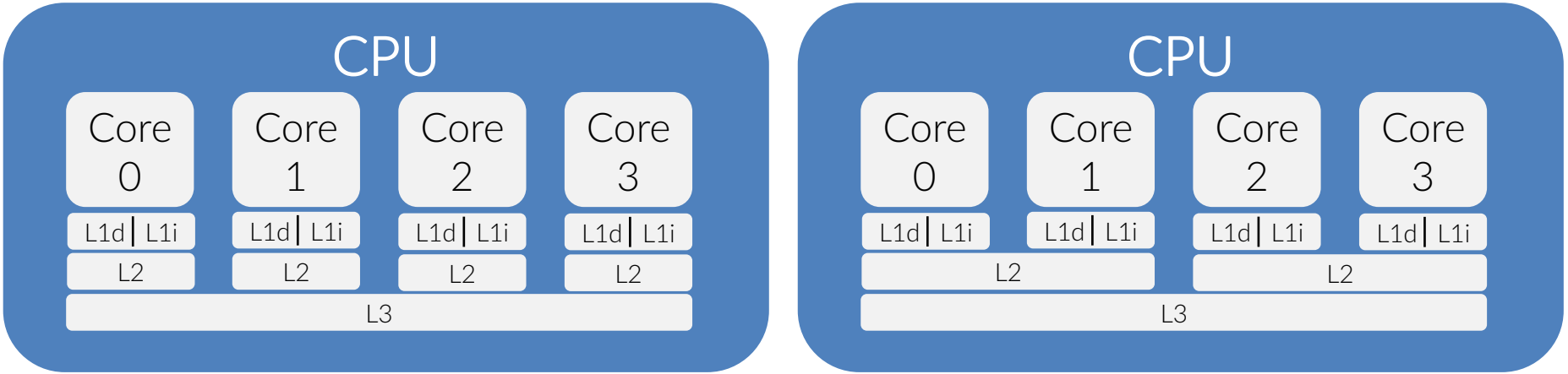
variant with shared memory and multiple channels;
coherency problem is alleviated

The typical NUMA architecture

- The bottleneck of resources (RAM and southbridge) access is eliminated
  - Each core *may* have its own DRAM module
- NUMA was originally developed to link several sockets, but it also evolved *inside* a single socket
- NEW problems:
  - data coherence (a variable *may* resides in a single DRAM module)
  - accessing DRAM has different costs

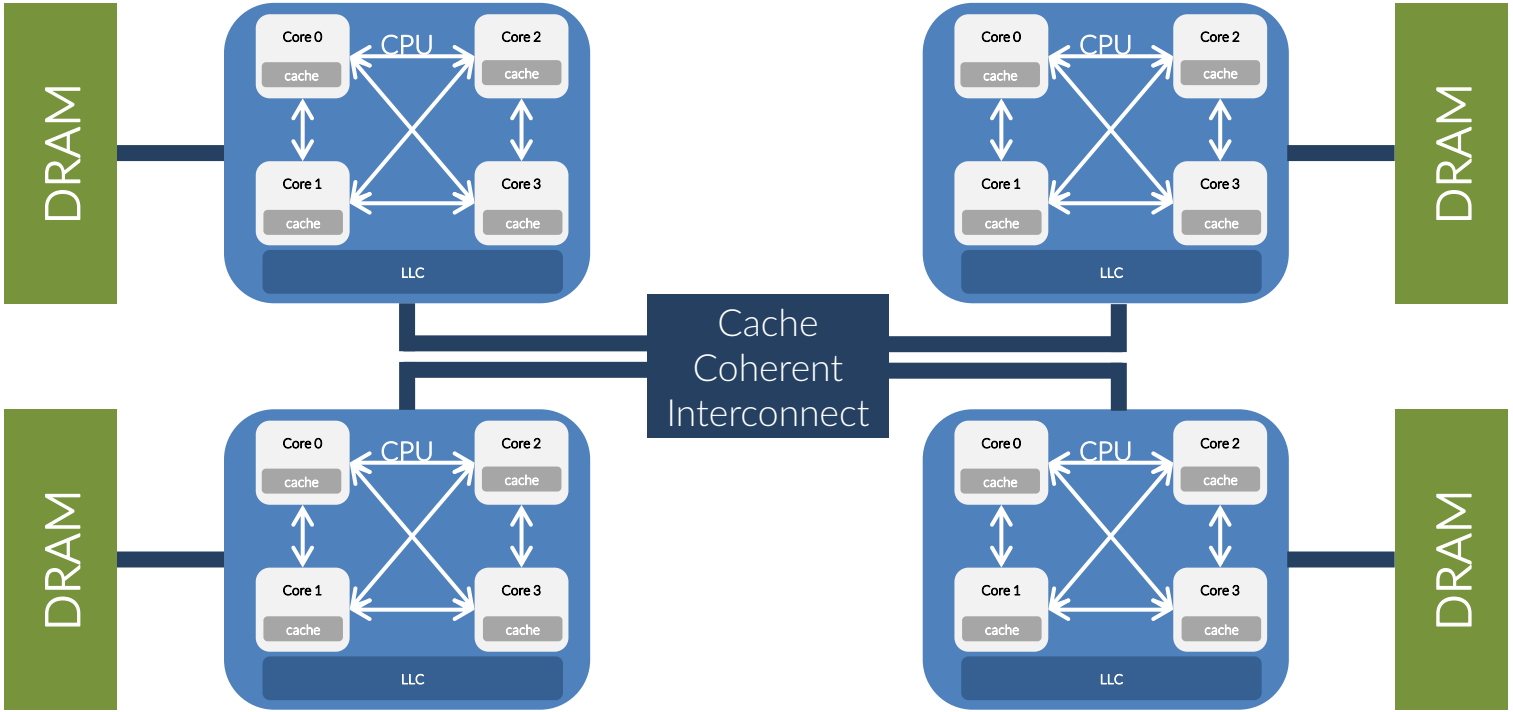
The typical NUMA architecture

Cache hierarchy can have different topologies



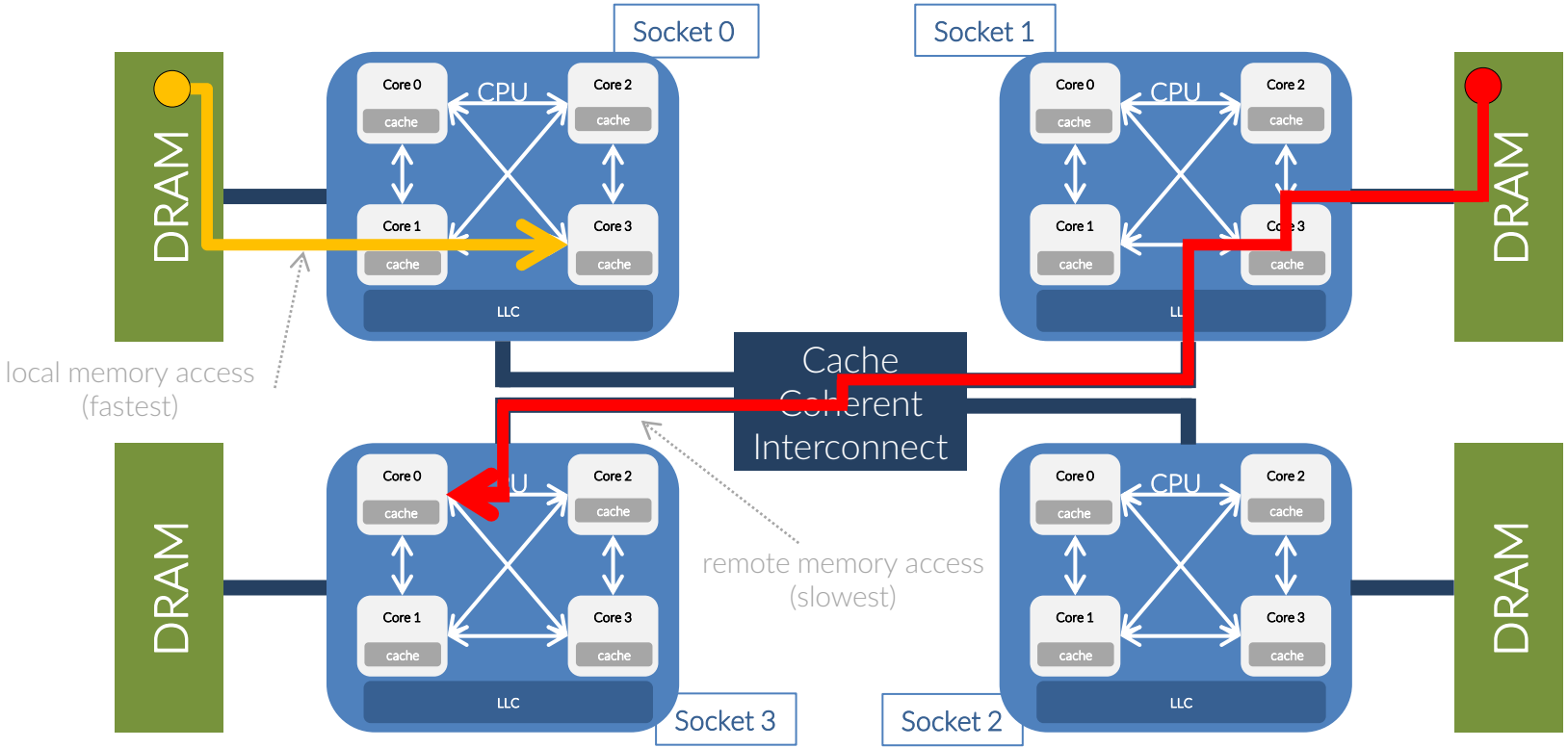
The typical NUMA architecture

A zoom-out to a multi-socket node



The typical NUMA architecture

Zoom-out to a multi-socket node (all the RAM is accessible from every core, i.e. it is shared)



The typical NUMA architecture

Two examples, both of nodes with 4 sockets each

```
[ltornatore@hp10 ~]$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                40
On-line CPU(s) list:   0-39
Thread(s) per core:    1
Core(s) per socket:    10
Socket(s):             4
NUMA node(s):          4
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 63
Model name:            Intel(R) Xeon(R) CPU E5-4627 v3 @ 2.60GHz
Stepping:              2
CPU MHz:               1200.000
CPU max MHz:           2600.0000
CPU min MHz:           1200.0000
BogoMIPS:              5194.05
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              256K
L3 cache:              25600K
NUMA node0 CPU(s):     0-4,20-24
NUMA node1 CPU(s):     5-9,25-29
NUMA node2 CPU(s):     10-14,30-34
NUMA node3 CPU(s):     15-19,35-39
```

hyperthreading off

```
node distances:
node   0   1   2   3
  0:  10  21  21  21
  1:  21  10  21  21
  2:  21  21  10  21
  3:  21  21  21  10
```

```
[ltornatore@gen10-01 ~]$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                96
On-line CPU(s) list:   0-95
Thread(s) per core:    2
Core(s) per socket:    12
Socket(s):             4
NUMA node(s):          4
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz
Stepping:              4
CPU MHz:               2663.305
CPU max MHz:           3200.0000
CPU min MHz:           1000.0000
BogoMIPS:              4600.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              16896K
NUMA node0 CPU(s):     0-11,48-59
NUMA node1 CPU(s):     12-23,60-71
NUMA node2 CPU(s):     24-35,72-83
NUMA node3 CPU(s):     36-47,84-95
```

hyperthreading on

```
node distances:
node   0   1   2   3
  0:  10  21  21  21
  1:  21  10  21  21
  2:  21  21  10  21
  3:  21  21  21  10
```

non-uniform access to memory

The typical NUMA architecture

How to discover the topology of your node:

- **`numactl`** tool
  it also controls the Linux NUMA policy
- **`cpuinfo`** tool (by Intel)
- **`hwloc`** (by OpenMPI)

The typical NUMA architecture

How to discover the topology of your node (examples):

- **`numactl`** tool
  `numactl -H`
- **`cpuinfo`** tool (by Intel)
- **`hwloc`** (by OpenMPI)

```
[ltornatore@hp08 ~]$ numactl -H
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 20 21 22 23 24
node 0 size: 65411 MB
node 0 free: 40998 MB
node 1 cpus: 5 6 7 8 9 25 26 27 28 29
node 1 size: 65536 MB
node 1 free: 58475 MB
node 2 cpus: 10 11 12 13 14 30 31 32 33 34
node 2 size: 65536 MB
node 2 free: 59344 MB
node 3 cpus: 15 16 17 18 19 35 36 37 38 39
node 3 size: 65536 MB
node 3 free: 59641 MB
node distances:
node   0   1   2   3
  0:  10  21  21  21
  1:  21  10  21  21
  2:  21  21  10  21
  3:  21  21  21  10
```

The typical NUMA architecture

How to discover the topology of your node (examples):

- **numactl** tool
- **cpuinfo** tool (by Intel)
- **hwloc** (by OpenMPI)

cpuinfo -d

```
=====  Placement on packages  =====
Package Id.     Core Id.        Processors
0               0,2,4,9,11,1,3,8,10,12          0,1,2,3,4,20,21,22,23,24
1               0,2,4,9,11,1,3,8,10,12          5,6,7,8,9,25,26,27,28,29
2               0,2,4,9,11,1,3,8,10,12          10,11,12,13,14,30,31,32,33,34
3               0,2,4,9,11,1,3,8,10,12          15,16,17,18,19,35,36,37,38,39
```

cpuinfo -g

```
=====  Processor composition  =====
Processor name    : Intel(R) Xeon(R)  E5-4627 v3
Packages(sockets) : 4
Cores             : 40
Processors(CPUs)  : 40
Cores per package : 10
Threads per core  : 1
```

cpuinfo -c

```
=====  Cache sharing  =====
Cache   Size            Processors
L1      32  KB          no sharing
L2      256 KB          no sharing
L3      25  MB          (0,1,2,3,4,20,21,22,23,24)(5,6,7,8,9,25,26,27,28,29)(10,11,12,13,14,30,31,32,33,34)(15,16,17,18,19,35,36,37,38,39)
```

The typical NUMA architecture

How to discover the topology of

- **numactl** tool
- **cpuinfo** tool (by Intel)
- **hwloc** (by OpenMPI)

```
Machine (256GB total)
  NUMANode L#0 (P#0 64GB)
    Package L#0 + L3 L#0 (25MB)
      L2 L#0 (256KB) + L1d L#0 (32KB) + L1i L#0 (32KB) + Core L#0 + PU L#0 (P#0)
      L2 L#1 (256KB) + L1d L#1 (32KB) + L1i L#1 (32KB) + Core L#1 + PU L#1 (P#1)
      L2 L#2 (256KB) + L1d L#2 (32KB) + L1i L#2 (32KB) + Core L#2 + PU L#2 (P#2)
      L2 L#3 (256KB) + L1d L#3 (32KB) + L1i L#3 (32KB) + Core L#3 + PU L#3 (P#3)
      L2 L#4 (256KB) + L1d L#4 (32KB) + L1i L#4 (32KB) + Core L#4 + PU L#4 (P#4)
      L2 L#5 (256KB) + L1d L#5 (32KB) + L1i L#5 (32KB) + Core L#5 + PU L#5 (P#20)
      L2 L#6 (256KB) + L1d L#6 (32KB) + L1i L#6 (32KB) + Core L#6 + PU L#6 (P#21)
      L2 L#7 (256KB) + L1d L#7 (32KB) + L1i L#7 (32KB) + Core L#7 + PU L#7 (P#22)
      L2 L#8 (256KB) + L1d L#8 (32KB) + L1i L#8 (32KB) + Core L#8 + PU L#8 (P#23)
      L2 L#9 (256KB) + L1d L#9 (32KB) + L1i L#9 (32KB) + Core L#9 + PU L#9 (P#24)
  NUMANode L#1 (P#1 64GB) + Package L#1 + L3 L#1 (25MB)
    L2 L#10 (256KB) + L1d L#10 (32KB) + L1i L#10 (32KB) + Core L#10 + PU L#10 (P#5)
    L2 L#11 (256KB) + L1d L#11 (32KB) + L1i L#11 (32KB) + Core L#11 + PU L#11 (P#6)
    L2 L#12 (256KB) + L1d L#12 (32KB) + L1i L#12 (32KB) + Core L#12 + PU L#12 (P#7)
    L2 L#13 (256KB) + L1d L#13 (32KB) + L1i L#13 (32KB) + Core L#13 + PU L#13 (P#8)
    L2 L#14 (256KB) + L1d L#14 (32KB) + L1i L#14 (32KB) + Core L#14 + PU L#14 (P#9)
    L2 L#15 (256KB) + L1d L#15 (32KB) + L1i L#15 (32KB) + Core L#15 + PU L#15 (P#25)
    L2 L#16 (256KB) + L1d L#16 (32KB) + L1i L#16 (32KB) + Core L#16 + PU L#16 (P#26)
    L2 L#17 (256KB) + L1d L#17 (32KB) + L1i L#17 (32KB) + Core L#17 + PU L#17 (P#27)
    L2 L#18 (256KB) + L1d L#18 (32KB) + L1i L#18 (32KB) + Core L#18 + PU L#18 (P#28)
    L2 L#19 (256KB) + L1d L#19 (32KB) + L1i L#19 (32KB) + Core L#19 + PU L#19 (P#29)
  NUMANode L#2 (P#2 64GB) + Package L#2 + L3 L#2 (25MB)
    L2 L#20 (256KB) + L1d L#20 (32KB) + L1i L#20 (32KB) + Core L#20 + PU L#20 (P#10)
    L2 L#21 (256KB) + L1d L#21 (32KB) + L1i L#21 (32KB) + Core L#21 + PU L#21 (P#11)
    L2 L#22 (256KB) + L1d L#22 (32KB) + L1i L#22 (32KB) + Core L#22 + PU L#22 (P#12)
    L2 L#23 (256KB) + L1d L#23 (32KB) + L1i L#23 (32KB) + Core L#23 + PU L#23 (P#13)
    L2 L#24 (256KB) + L1d L#24 (32KB) + L1i L#24 (32KB) + Core L#24 + PU L#24 (P#14)
    L2 L#25 (256KB) + L1d L#25 (32KB) + L1i L#25 (32KB) + Core L#25 + PU L#25 (P#30)
    L2 L#26 (256KB) + L1d L#26 (32KB) + L1i L#26 (32KB) + Core L#26 + PU L#26 (P#31)
    L2 L#27 (256KB) + L1d L#27 (32KB) + L1i L#27 (32KB) + Core L#27 + PU L#27 (P#32)
    L2 L#28 (256KB) + L1d L#28 (32KB) + L1i L#28 (32KB) + Core L#28 + PU L#28 (P#33)
    L2 L#29 (256KB) + L1d L#29 (32KB) + L1i L#29 (32KB) + Core L#29 + PU L#29 (P#34)
  NUMANode L#3 (P#3 64GB) + Package L#3 + L3 L#3 (25MB)
    L2 L#30 (256KB) + L1d L#30 (32KB) + L1i L#30 (32KB) + Core L#30 + PU L#30 (P#15)
    L2 L#31 (256KB) + L1d L#31 (32KB) + L1i L#31 (32KB) + Core L#31 + PU L#31 (P#16)
    L2 L#32 (256KB) + L1d L#32 (32KB) + L1i L#32 (32KB) + Core L#32 + PU L#32 (P#17)
    L2 L#33 (256KB) + L1d L#33 (32KB) + L1i L#33 (32KB) + Core L#33 + PU L#33 (P#18)
    L2 L#34 (256KB) + L1d L#34 (32KB) + L1i L#34 (32KB) + Core L#34 + PU L#34 (P#19)
    L2 L#35 (256KB) + L1d L#35 (32KB) + L1i L#35 (32KB) + Core L#35 + PU L#35 (P#35)
    L2 L#36 (256KB) + L1d L#36 (32KB) + L1i L#36 (32KB) + Core L#36 + PU L#36 (P#36)
    L2 L#37 (256KB) + L1d L#37 (32KB) + L1i L#37 (32KB) + Core L#37 + PU L#37 (P#37)
    L2 L#38 (256KB) + L1d L#38 (32KB) + L1i L#38 (32KB) + Core L#38 + PU L#38 (P#38)
    L2 L#39 (256KB) + L1d L#39 (32KB) + L1i L#39 (32KB) + Core L#39 + PU L#39 (P#39)
```

The typical NUMA architecture

How to discover the topology of your node:

- **`numactl`** tool
- **`cpuinfo`** tool (by Intel)
- **`hwloc`** (by OpenMPI)
- directly exploring **`/sys/devices/system/`**

The typical NUMA architecture

How to discover the topology of your node:

- `numactl` to
- `cpuinfo` to
- `hwloc` (by C
- directly exp `tem/`

Jump to Numa Awareness to explore how to orchestrate data and OpenMP threads on NUMA architectures

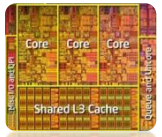
Cache coherence

Data synchronization is one of the main performance killers for multi-core applications.

- when a memory region is accessed by two cores (i.e. by two different threads running on two different cores), it must be present in both L1/L2, and when one core updates the value, the change must be propagate.

- when a thread migrates, the data will still resides on another's core memory.

Memory consistency for the whole system is guaranteed at hardware level, resulting in huge wasting of time if data are not properly handled.
For instance, concurrent access in writing is a main sink of cpu cycles.



Cache coherence: MESI

Data consistency is maintained by the **MESI** standard.

It is the successor of the MSI protocol and the ancestor of MESOI one

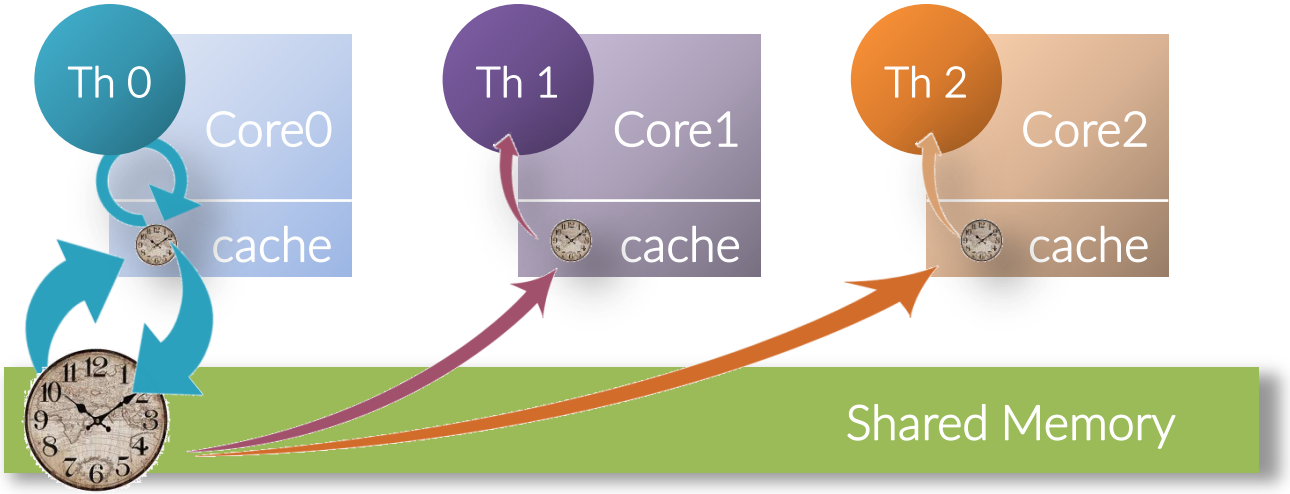
see:
MD64 Architecture Programmer's Manual
Volume 2: System Programming

MODIFIED	*X*'s values has been modified by this core, and then this is the only valid copy in the system
EXCLUSIVE	*X* is used by this core only; changes do not need to be signalled
SHARED	*X* is used by multiple cores; changes need to be signalled
INVALID	*X*'s value has been modified by another core (or *X* is not used)

"X" stands for a given memory location

Cache coherence: MESI – an example

Let's clarify with an example. Let's say that there are 3 threads, running on separate cores, accessing some shared-memory.
`Thread0` is running the application `clock()`, which ticks a shared-memory variable that contains the wall-clock `time`. In time to time, both `Thread1` and `Thread2` want to know what `time` it is.



Cache coherence: MESI – an example

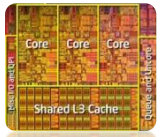
		Time (in secs)	Action	cache status		
M	Modified			Core0	Core1	Core2
E	Exclusive	0	-	I	I	I
S	Shared	1	Th0 reads	E	I	I
I	Invalid	2	Th0 writes[0]	E	I	I
		2.3	Th1 reads[1]	S	S	I
		2.7	Th2 reads	S	S	S
		3	Th0 writes[2]	M	I	I
		4	Th0 writes	M	I	I
		4.4	Th2 reads[1]	S	S	S
		5	Th0 writes	M	I	I
		…	…	…	…	…

[0] Core0 is the only one using the value, that is then "Exclusive". No signal needs to be sent around.

[1] A signal is issued to "the memory", which recognizes that the only valid copy is in the Core0 cache.
Hence, that value is copied back into the shared memory, and from there it is copied in the cache. At that point, everybody has a valid copy, which is then "Shared" (*).

[2] A signal about the change is issued to all the interested actors (those who have a copy) because their values are now "Invalid".
0's copy is instead "Modified".

(*) In the MESOI protocol, in this case the copy can be sent directly to the other caches, without having to transit by the DRAM



Great powers, great responsibility

A — **Variables used by a single core**
They should resides in a single cache

B — **Read-only variables**
No issues in being shared among many cores

C — **Modified variables**
Variables modified by many cores, or read by many cores

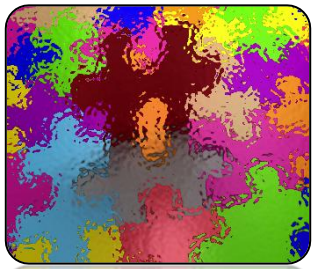
A cache line should contain only one type of data.

Put variables in the order they will be used.

Modified variables should stay together, they are the bottlenecks.

The **false sharing** happens when variables of type A or B resides in the same cache line of a type C. Or when two type C variables, modified by two different cores, reside in the same cache line.

Introduction Outline



Intro to Parallel Computing

Parallel Performance

What is parallel computing ?

1. A **parallel computer** is a computational system that offers *simultaneous access* to *many computational units* fed by memory units.
   The computational units are required to be able to *co-operate* in some way, meaning *exchanging data and instructions* with the other computational units.

2. **Parallel processing** is the ensemble of techniques and algorithms that makes you able to actually use a parallel computer to successfully and *efficiently* solve a problem.

What is parallel computing ?

The parallel processing is expressed by **software entities** that have an increasing level of granularity:

processes, threads, routines, loops, instructions..

The software entities run on underlying **computational hardware entities** as

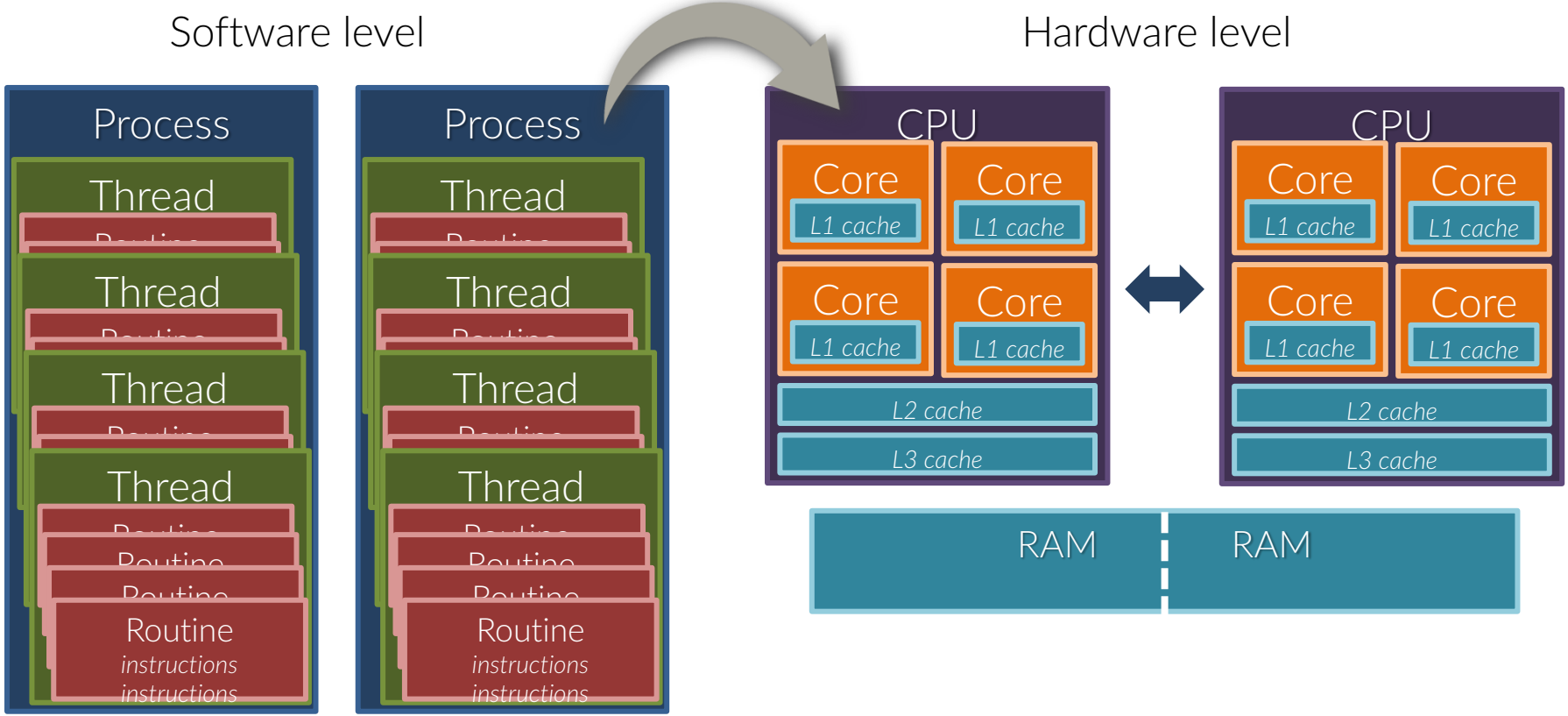
processors, cores, accelerators

The data to be processed/created live and travel in **storage hardware entities** as

Memory, caches, NVM, networks, DMA

The *exploitation/access* of hardware resources (computational and storage) is **concurrent** among software entities

What is parallel computing ?



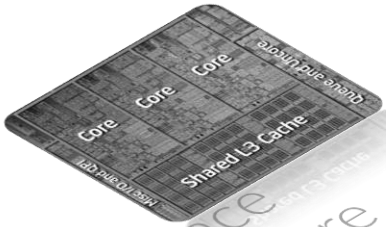
For two main reasons:

1. Time-to-solution
   - To solve the same problem in a smaller time
   - To solve a larger problem in the same time

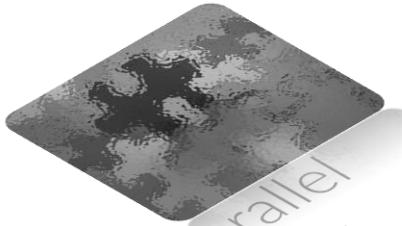
2. Problem size (~ data size)

   To solve a problem that could *not* fit on the memory addressable by a single computational units (or that could fit in the space around a single computational units without serious performance loss)

Introduction Outline



The race to multicore



Intro to Parallel Computing

Parallel Performance

What is parallel performance

Has we have seen, «performance» is a tag that can stand for many things.

In this frame, with «performance» we mean the relation between the computational requirements and the computational resources needed to meet those requirements.

Resources

- Hardware
- Time
- Energy
- Money

$$\text{Performance} \approx \frac{1}{resources}$$

$$\text{Performance ratios} \approx \frac{resources_1}{resources_2}$$

Inspired by "parallel performance theory, CIS 410/510 Univ. of Oregon

What is parallel performance

Performance is a measure of how well the computational requirements are met and, at the same time, of how well the computational resources are exploited.

"The most constant difficulty in contriving the engine has arisen from the desire to reduce the time in which the calculations were executed to the shortest which is possible."
Charles Babbage, 1791 – 1871

Key factors

`n`	Problem size
$T_s(n)$	Serial run-time
$T_p(n)$	Parallel run-time
`p`	Number of computing units
f_n	Intrinsic sequential fraction of the problem of size n
`k(n, t)`	Parallel overhead

Speedup $$Sp(n,t) = \frac{T_s(n)}{T_p(n)}$$

Efficiency $$Eff(n,t) = \frac{T_s(n)}{p \times T_p(n)} = \frac{Sp(n,t)}{p}$$

Naïve expectations

- If single processor ~ *m Mflops*, parallel flops performance with *p* tasks is $p \times m$ *Mflops*.
- If sequential run-time is T, parallel run-time with p tasks is $\propto T/p$.
- If parallel run-time with p tasks is T, and the run-time with p_1 tasks is T_1, then $T_1/T_2 \propto p_2/p_1$
- If parallel run-time with p tasks and problem size Z is T, the run-time with size Z_1 is $T_1 \propto T \times Z_1/Z$.

Is that correct ?

Parallel performance

Sequential execution time is

$T_S = T(n,1) \times f_n + T(n,1) \times (1-f_n)$

Assuming that the parallel fraction of the computation is *perfectly parallel*, parallel execution time is

$T_P = T(n,p) = T_S \times f_n + T_S \times (1-f_n)/p + k(n,t)$

And then

speedup $= Sp(n,p) \leq T_S / T_P$

efficiency $= Eff(n,p) \leq Sp(n,p) / p$

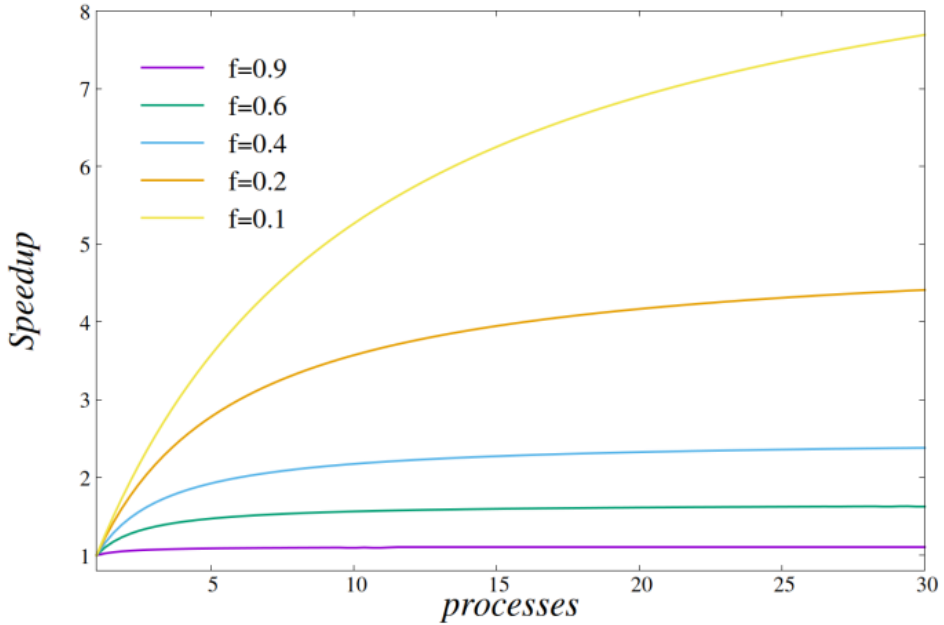
Amdahl's law

If f is the fraction of the code which is intrinsically sequential,

the speedup is then

$$Sp(n,t) \leq \frac{1}{f + \frac{1-f}{p}} \leq \frac{1}{f}$$

Note that we wrote f instead of f_n



Amdahl's law

There are some significant issues in the Amdhal's law shown in the previous slide:

- No matter of how many processes *p* are used, the speedup is determined by *f* (and is quite low for ordinary problems).
- The problem size *n* is kept fixed when estimating the possible speedup while the number of processes increases (*strong scaling*).
  However, most often the problem size increases as well.
- The parallel overhead *k(n, t)* is ignored, which leads to an optimistic estimate of the speedup, and usually, $p(n)/t > k(n,t)$
- The fraction of sequential part may decrease when the problem size increases

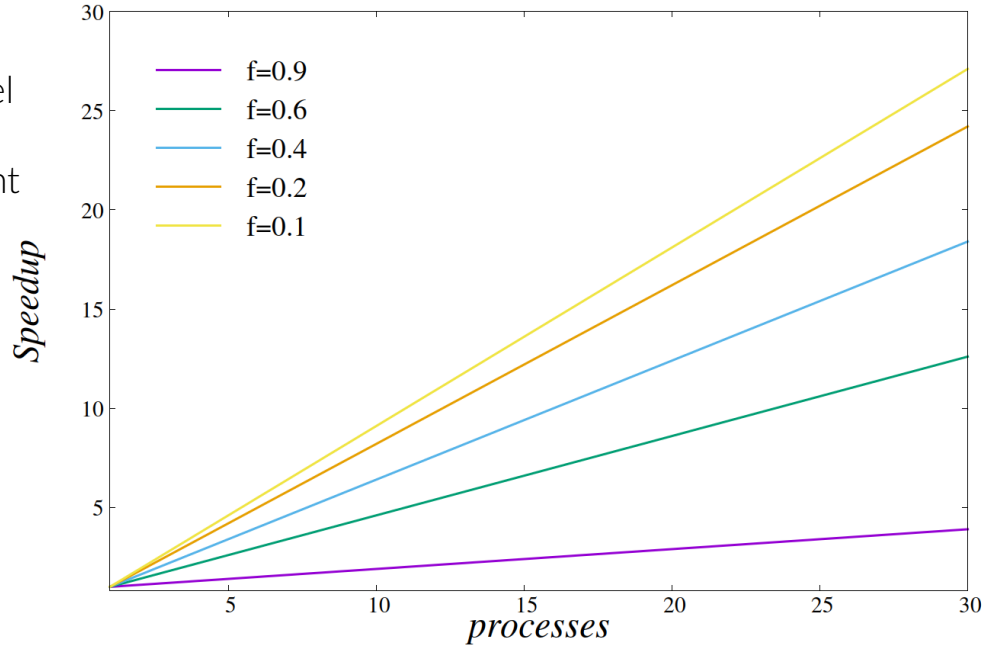
Then, usually the speedup increases with problem size

Gustafson's law

However, normally when you increase the problem's size, the parallelizable part increases way more than the sequential part.

If we consider the workload as the sum
$w = a+b$
where a and b being the serial and the parallel work, and we assign the same amount of workload to every process, that would amount to a serial run-time
$T_S \propto a + p \times b$
while it still takes
$T_P \propto a+b$ using p processes
Hence the speedup is $[a + p \times b] / [a+b]$, which if $f_n = a/(a+b)$ we can rewrite as the Gustafson's law for the speedup:

$$Sp_G(n,t) = p - (p-1)f_n \leq p$$



Scalability

The two lines of reasoning, the former by Amdhal and the latter by Gustafson, lead us to two different concepts for the *scalability*, which is the ability of a parallel system to increase its efficiency when the number of processes and/or the size of the problem get larger.

1. STRONG SCALABILITY: the problem size is fixed, p increases

2. WEAK SCALABILITY: the workload is fixed, the problem size *and* p increase

Parallel overhead

In parallel computing they may be several sources of overhead due to the parallelization itself:

- Communication overhead
- Algorithmic overhead
- Synchronization
  - Critical paths - Dependencies across different processes
  - Bottlenecks (some processes are stuck and make all the others to waist time)
  - Work-load imbalance
- Thread/processes creation

Hence, if *k(n,t)* is the overhead of some kind, t_S *and* t_P the run-time for the serial and the parallel part, the parallel run-time can be written as

$$T_P(n,p) = t_S + \frac{t_P}{p} + k(n,t)$$

Let's define an experimentally measured serial fraction of time:

$$e(n,p) = \frac{t_S + k(n,p)}{t_S + t_P}$$

Parallel overhead

With a little bit of math: $$e(n,p) = \frac{\frac{1}{Sp(n,t)} - \frac{1}{p}}{1 - \frac{1}{p}}$$

Let's check a couple of examples:

p	2	4	8	10	20
Sp(p)	1.69	2.6	3.52	3.79	4.49
E(p)	0.18	0.18	0.18	0.18	0.18

The measured serial fraction is constant, the lack of scaling is due to the 0.18 fraction of serial workload.

p	2	4	8	10	20
Sp(p)	1.67	2.47	3.11	3.22	3.15
E(p)	0.2	0.21	0.22	0.23	0.28

The measured serial fraction keep increasing: the lack of scaling is also due parallelization overhead

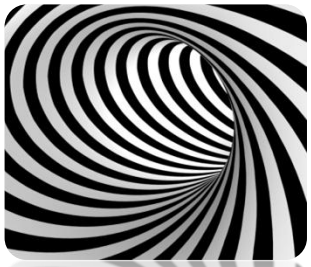
Outline

Intro to OpenMP

OpenMP Outline

Introduction & Concept

Parallel Regions



Parallel Loops

NUMA AWARENESS

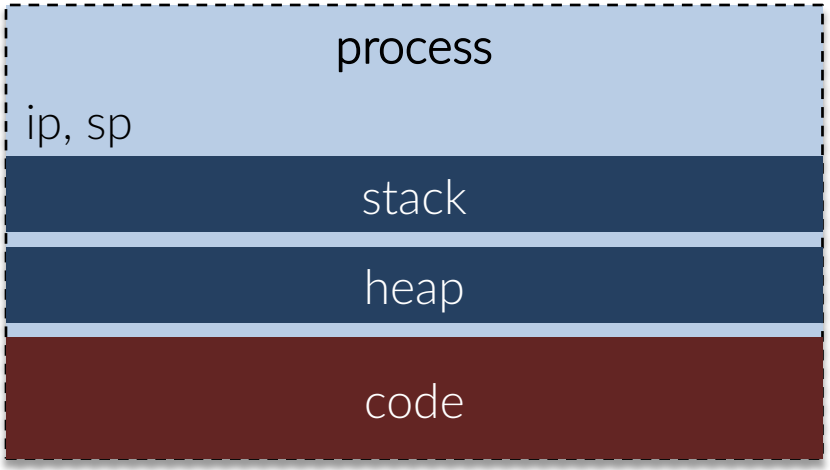
Sections & Task

Threads and processes

A **process** is an independent sequence of instructions *and* the ensemble of resources needed for their execution.
A program needs much more than just its binary code (i.e. the list of ops to be executed): it needs to access to a protected memory space and to access system resources (e.g. files and network).

A "process" is then a program that has been allocated with the necessary resources by the operating system.

There may be different **instances** of the same program as different, independent processes

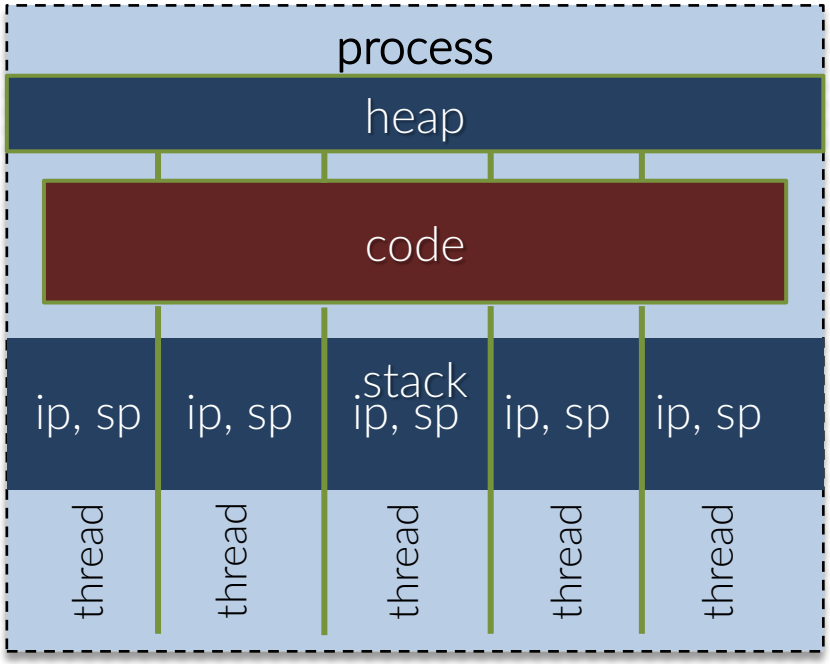


Threads and processes

A **thread** is an independent instance of code execution *within* a process. There may be from one to many threads within the same process.

Each thread shares the same code, memory address space and resources than its father process.
While each thread has its own stack, ip and sp, the heap will be shared among threads, which then operate in *shared-memory*.

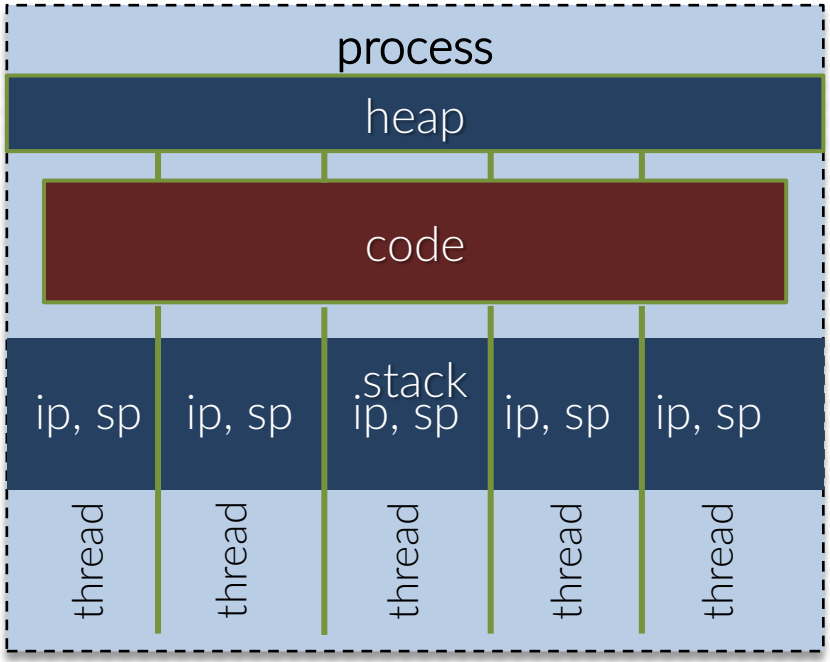
Spawning threads inside a process is much less costly than creating processes.



Threads and processes

A thread can run either on the same computational units of its father process or on a different one.

A computational unit nowadays amounts to a **core**, either inside the same CPU (socket) on which the father process runs, or inside a sibling socket in the same NUMA region.



What is OpenMP

OpenMP is a standard API to enable shared-memory parallel programming: **Open** specifications for **M**ulti**P**rocessing

It allows to write multi-threaded programs with a standard behaviour through the usage of a set of compiler directives to be inserted in the source code:

- Pragmas '#' in C/C++
- Specially formatted comments in Fortran

Both fine- and coarse-grain parallelism are possible, from loop-level to explicit assignment to threads.

What is OpenMP

1997	OpenMP for Fortran, 1.0
1998	OpenMP for C/C++, 1.0
2000	OpenMP for Fortran, 2.0
2002	OpenMP for C/C++, 2.0
2005	OpenMP 2.5
2008	OpenMP 3.0
2013	OpenMP 4.0
2015	OpenMP 4.5
2016	OpenMP 5.0

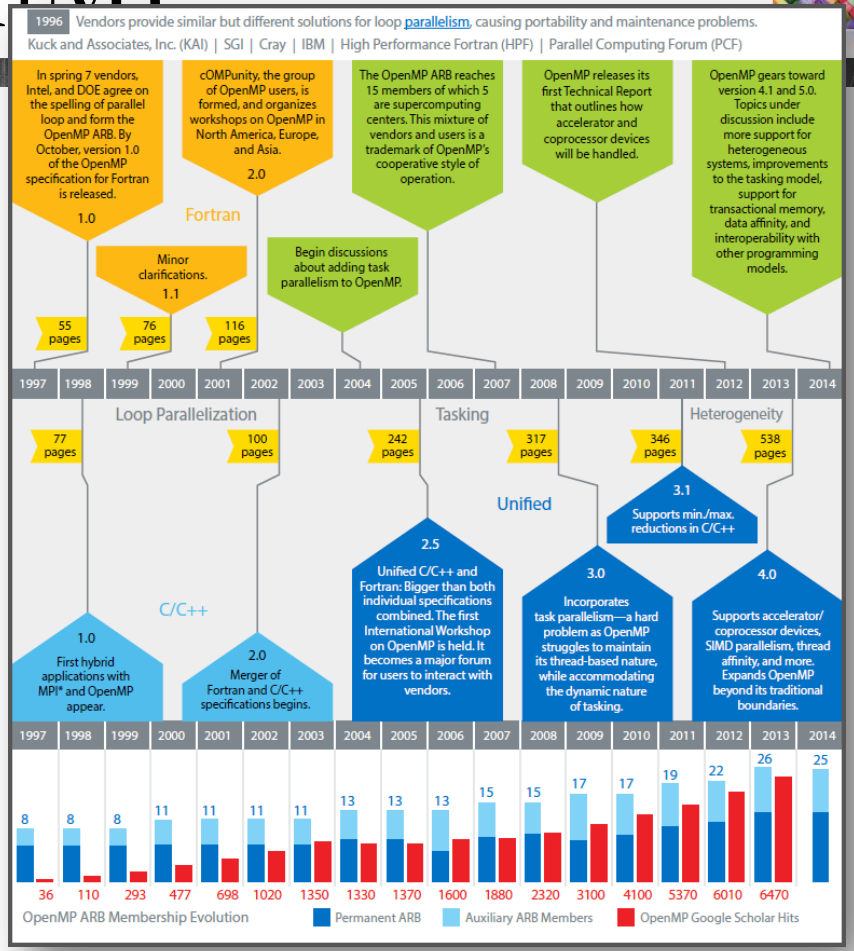
regular loops parallelization

Irregular parallelism task-based parallelism

accelerators / tasks+ / atomics / affinity / SIMD / user reduction

off-loading to accelerators; more on tasks and workshares, ...

from http://openmp.org



What is OpenMP

Advantages of a directive-based approach

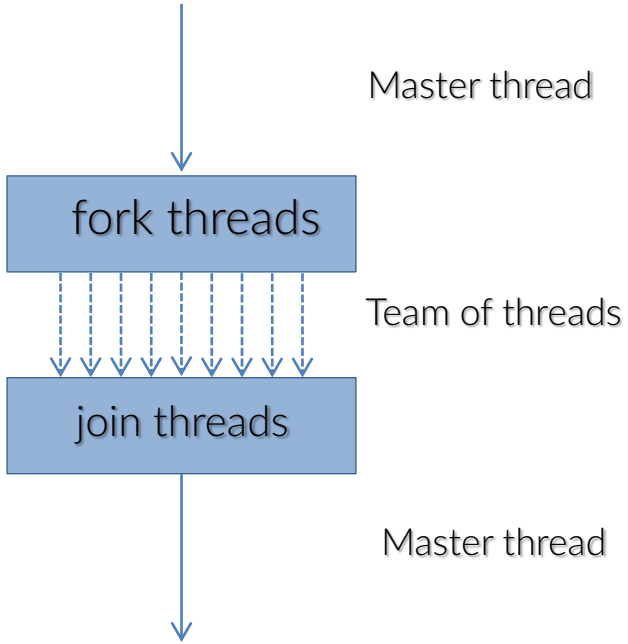
- **Abstraction**
  Subtleties of `pthread` and hardware-specific aspects are hidden. You can focus on data and workflow much more easily.
- **Efficiency.**
  The learning curve to achieve reasonable results is much shallower. The code's design is easier, the result/effort ratio is favourable with respect to `pthread`.
- **Incremental approach**
  No need to re-write your whole code. You start concentrating on some sections only, following a the suggestions from profiling.
- **Portability.**
  The compiler will take care of this for you. You still have to develop a design able to adapt to different topologies.
- **One source**
  Through conditional compilation, serial and parallel versions can easily coexist.

OpenMP programming model

use a single master thread for serial operations

spawn a team of threads to perform parallel work

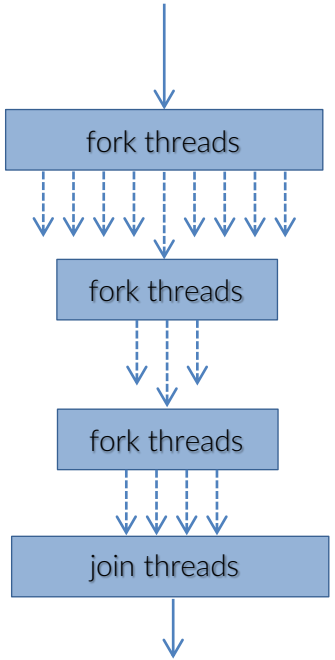
use a single master thread for serial operations



OpenMP programming model

- Threads access and modify shared memory regions
  - explicit or implicit synchronization protect against race conditions
  - there is no concept like explicit “message-passing”
  - loop-carried dependencies hamper any parallel speedup
  - shared-variable attributes are vital to reduce or avoid race conditions or the need for synchronization
- Each thread performs its part of parallel work in a separate space and stack that are not visible to other threads or outside the parallel region
- Nested parallelism is explicitly permitted
- The number of threads can be dynamically changed before a parallel region

NESTED PARALLELISM



OpenMP directives

An OpenMP directive is a specially-formatted pragma for C/C++ and comment for FORTRAN codes.
Most of the directives apply to *structured code block*, i.e. a block with a single input and a single output points and no branch within it.

The directives allows to

- create team of threads for parallel execution
- manage the sharing of workload among threads
- specify which memory regions (i.e. variables) are shared and which are private to each threads
- drive the update of shared memory regions
- synchronize threads and determine atomic/exclusive operations

DECLARE PARALLEL REGION

```
!$OMP PARALLEL
  …
!$OMP END PARALLEL
```

```
#pragma omp parallel
{
  …
}
```

Dynamic extent

As we have seen in the previous slide, the lexical scope of structured blocks defines the *static* extent of an OpenMP parallel region.

Every function call from within a parallel region determines the creation of a *dynamic* extent to which the same directives apply.
The dynamic extent includes the original static extent and all the instructions and further calls along the call tree.
The functions called in the dynamic extent can contain additional OpenMP directives.

```
#pragma omp parallel
{
   double *array;
   int N;
   …
   sum = foo(array, N);
   …
}

double foo( double *A, int N )
{
   double sum = 0;
   #pragma parallel for reduction(+:sum)
   for ( int ii = 0; ii < N; ii++ )
      sum += array[ii];
   return sum;
 }
```

static extent

dynamic extent

"orphan" directive

OpenMP toolbox

OpenMP is made of 3 components:

1. Compiler directives
   give indication to the compiler about how to manage threads internals

2. Run-time libraries linked by the compiler

3. Environment variables
   set by the user, determine the behaviour of the omp library; for instance, the number of threads to be spawned or the requirements about the thread-cores-memory affinity

OpenMP toolbox

- directives for parallel regions
  - #pragma omp parallel directive
- work-sharing constructs
  - #pragma omp for
  - #pragma omp section
  - #pragma omp task
- data attributes
  - shared, private, firstprivate, lastprivate, copyin, copyprivate, threadprivate clauses
- Synchronization
  - critical, atomic, barrier, ordered, flush, nowait
- run-time
  - Lot of run-time functions and utils
    omp_set_num_threads(), omp_get_thread_num(), omp_get_num_threads(), …

Conditional compilation

By default, when the compiler is instructed to activate the processing of OpenMP directives,

```
gcc  -fopenmp …
icc  -fopenmp …
pgcc -mp …
```

it defines a macro that let you to conditionally compile sections of the code:

```
   …
#ifdef _OPENMP
   …
   …
#endif
   …
```

omp_versions.c : with this example, you see how to determine what version of OpenMP standard is supported by your compiler

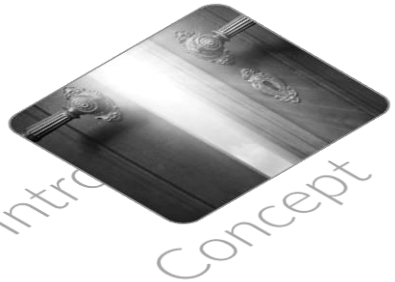
Conditional compilation

What is this useful for?

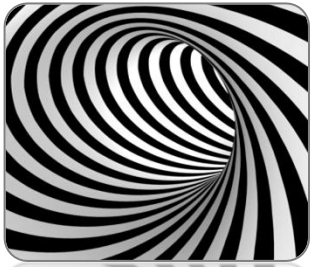
To write code that works as well also without OpenMP.

That helps you in assessing the correctness and portability of your code (mostly if you are writing an hybrid code, for instance MPI+OpenMP).

OpenMP Outline



Parallel Regions



Parallel Loops

Advanced Parallelism

NUMA AWARENESS

Parallel Regions

```
#pragma omp parallel _some_clauses_here_
single-line-here
```

A parallel region can be as short as a single line

```
#pragma omp parallel _some_clauses_here_
{ … }
```

There are no limits on the size of the code included within {..}.

```
#pragma omp parallel for _some_clauses_here_
{ … }
```

The specific construct about `for` loops

```
#pragma omp sections _some_clauses_here_
{ … }
```

A more general work-sharing construct

```
#pragma omp task _some_clauses_here_
{ … }
```

This allows task-based parallelism

The region starts at the opening { brace and ends at the closing } one.
An implicit synchronization barrier is present at the end of the region (*).

However, for efficiency reasons, it may be that the threads do not die at the end of the region but remain alive until the father process dies

(*) we'll see the details about synchronization later on

The threads factory

When you create a parallel region, through an OpenMP directive, a pool of threads is created.
Each one receives an ID, ranging from 0 to *n-1*, where *n* is the number of threads.

Their stack and IP are separated from the others' ones and from the father-process' ones.

Can we check that?
What is the fate of the creating process (thread) ?

```
int    i;

register unsigned long long base_of_stack asm("rbp");
register unsigned long long top_of_stack asm("rsp");

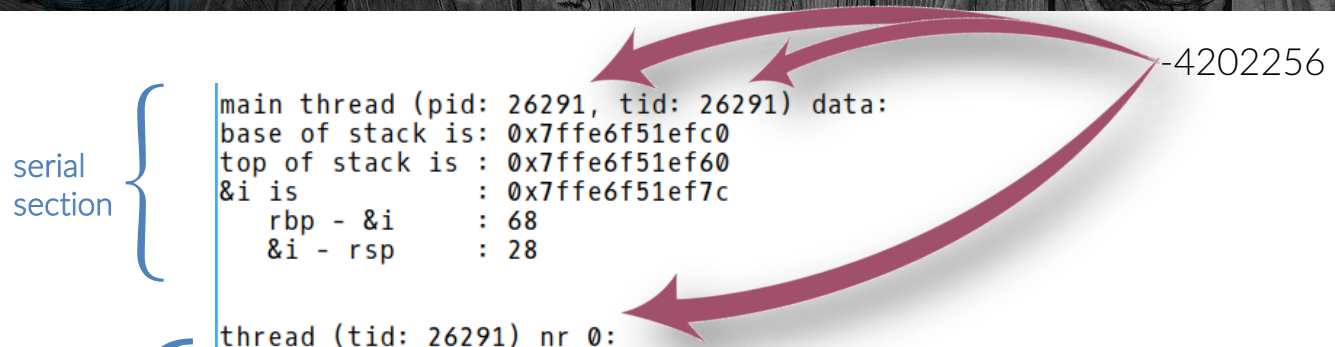
 printf( "\nmain thread (pid: %d, tid: %d) data:\n"
         "base of stack is: %p\n"
         "top of stack is : %p\n"
         "&i is           : %p\n"
         "   rbp - &i     : %td\n"
         "   &i - rsp     : %td\n"
         "\n\n",
         getpid(), syscall(SYS_gettid),
         (void*)base_of_stack,
         (void*)top_of_stack,
         &i,
         (void*)base_of_stack - (void*)&i,
         (void*)&i - (void*)top_of_stack );

#pragma omp parallel private(i)
  {
    int me = omp_get_thread_num();
    unsigned long long my_stackbase;
    __asm__("mov %%rbp,%0" : "=mr" (my_stackbase));

    printf( "\tthread (tid: %ld) nr %d:\n"
                      "\t\tmy base of stack is %p ( %td from main\'s stack )\n",
                      "\t\tmy i address is %p\n"
                      "\t\t\t%td from my stackbase and %td from main\'s\n",
                      syscall(SYS_gettid), me,
                      (void*)my_stackbase, (void*)base_of_stack - (void*)my_stackbase,
                      &i, (void*)&i - (void*)my_stackbase,
                      (void*)&i - (void*)base_of_stack);
}
```

parallel_regions/
00_stack_and_scope.c

The threads factory



serial section

```
main thread (pid: 26291, tid: 26291) data:
base of stack is: 0x7ffe6f51efc0
top of stack is : 0x7ffe6f51ef60
&i is           : 0x7ffe6f51ef7c
   rbp - &i     : 68
   &i - rsp     : 28
```

parallel region

```
thread (tid: 26291) nr 0:
        my base of stack is 0x7ffe6f51eee0 ( 224 from main's stack )
        my i address is 0x7ffe6f51eea0
                -64 from my stackbase and -288 from main's
thread (tid: 26293) nr 2:
        my base of stack is 0x7f7342c82ba0 ( 597747680288 from main's stack )
        my i address is 0x7f7342c82b60
                -64 from my stackbase and -597747680352 from main's
thread (tid: 26294) nr 3:
        my base of stack is 0x7f7342880ba0 ( 597751882784 from main's stack )
        my i address is 0x7f7342880b60
                -64 from my stackbase and -597751882848 from main's
thread (tid: 26292) nr 1:
        my base of stack is 0x7f7343084b20 ( 597743477920 from main's stack )
        my i address is 0x7f7343084ae0
                -64 from my stackbase and -597743477984 from main's
```

- Each thread has its own stack; `thread 0` inherits its own stack, which has grown to host the data relative to OpenMP parallel region;
- the private `i`s are actually in the threads stacks, and so different memory regions than the shared `i`.

`parallel_regions/00_stack_and_scope.c`

The threads factory

How large is the stack of each thread? The default value is system dependent.
That is the output of the `00_stack_and_scope` code example: it prints the `BP` and `SP` pointers of each thread. Then, we know how large is the threads' stack at a given moment (only an integer is defined, plus the system's thread structure), and how much memory is reserved for the stack to grow:

```
thread 0: bp @ 0x7ffc6edea2f0, tp @ 0x7ffc6edea280: 112 B
thread 1: bp @ 0x7fa371d18b20, tp @ 0x7fa371d18ab0: 112 B        --> -382202615648 B from top of thread 0
thread 2: bp @ 0x7fa371916ba0, tp @ 0x7fa371916b30: 112 B        --> -4202256 B from top of thread 1
thread 3: bp @ 0x7fa371514ba0, tp @ 0x7fa371514b30: 112 B        --> -4202384 B from top of thread 2

you did not define OMP_STACKSIZE: try to set it and check what happens to the threads' stack pointers
```

In this case, compiled with gcc, it seems that about 4MB are reserved for each thread. The same value holds with pgi, whereas clang seems to reserve 132MB (i.e., the addresses in the virtual space are spaced by ì132MB).
However, you can control the size of the threads' stack using the environmental variable `OMP_STACKSIZE`:

```
export OMP_STACKSIZE=N
```

Where `N` is a number, followed by a letter: '`B`', '`K`', '`M`', '`G`' mean bytes, kilobytes, megabytes and gigabytes respectively (if no letter is specified, the number is interpreted as kilobytes).

The threads factory


```
export OMP_STACKSIZE=32K
```

```
thread 0: bp @ 0x7ffd1a2bbc70, tp @ 0x7ffd1a2bbbd0: 160 B
thread 1: bp @ 0x7f18b4be4af0, tp @ 0x7f18b4be4a50: 160 B        --> -980954214624 B from top of thread 0
thread 2: bp @ 0x7f18b4bdab70, tp @ 0x7f18b4bdaad0: 160 B        --> -40672 B from top of thread 1
thread 3: bp @ 0x7f18b4bd0b70, tp @ 0x7f18b4bd0ad0: 160 B        --> -40800 B from top of thread 2

you defined OMP_STACKSIZE as 32K: try to change that value and check what happens to the threads' stack pointers
```

```
export OMP_STACKSIZE=1M
```

```
thread 0: bp @ 0x7ffc6f8b97f0, tp @ 0x7ffc6f8b9750: 160 B
thread 1: bp @ 0x7fd5bf930af0, tp @ 0x7fd5bf930a50: 160 B        --> -166161058912 B from top of thread 0
thread 2: bp @ 0x7fd5bec96b70, tp @ 0x7fd5bec96ad0: 160 B        --> -13213408 B from top of thread 1
thread 3: bp @ 0x7fd5beb94b70, tp @ 0x7fd5beb94ad0: 160 B        --> -1056608 B from top of thread 2

you defined OMP_STACKSIZE as 1M: try to change that value and check what happens to the threads' stack pointers
```

The threads factory

How many thread can be created by a single process?
Also this limit is system-dependent.

On Linux, it depends on the amount of total physical memory: basically, the maximum number of threads is the amount of physical memory divided by the memory amount needed to describe and run a thread, times a factor that accounts for the fact that you don't want all the memory allocated only to make the threads alive.

You can check this limit by

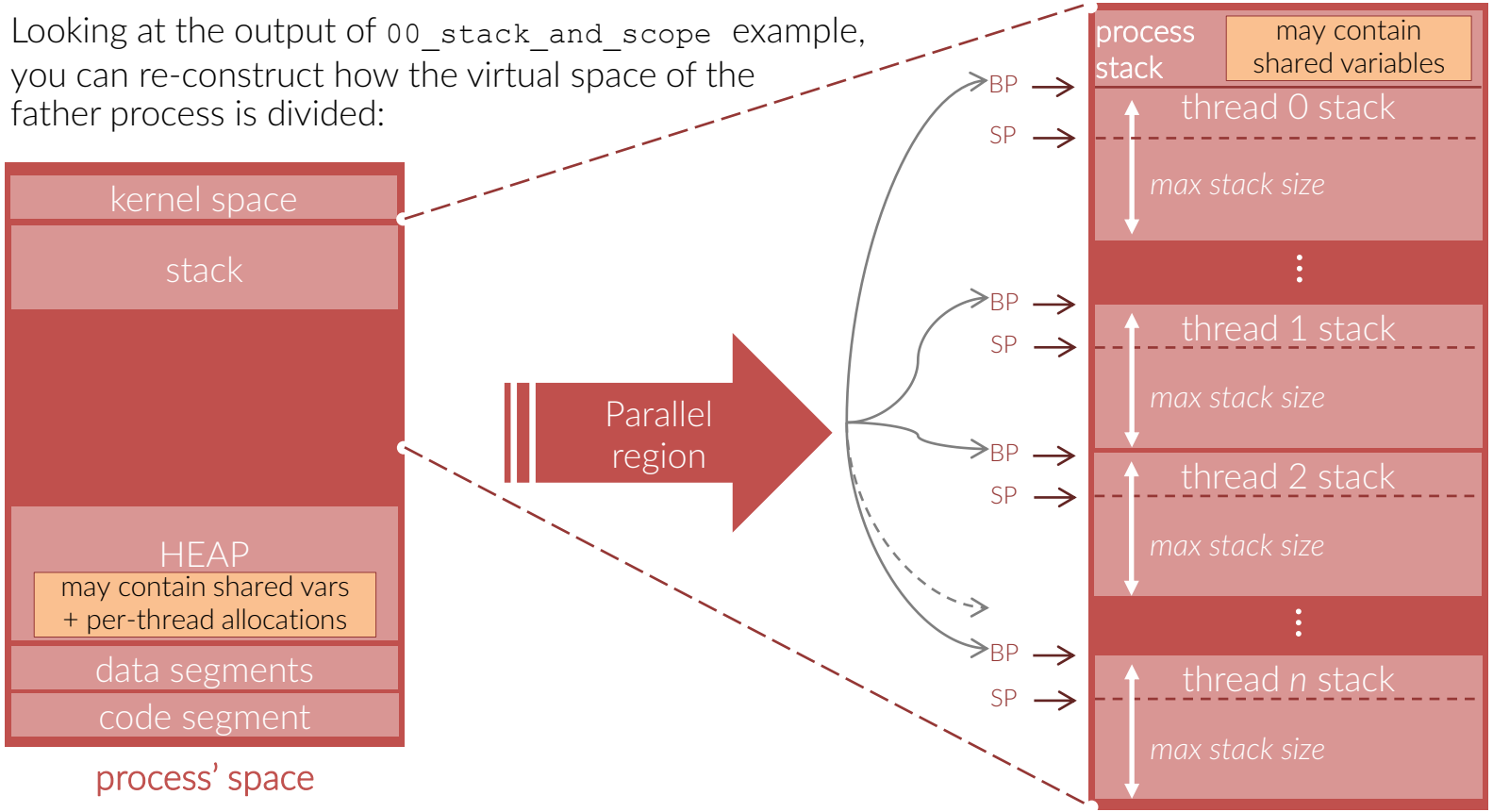
```
cat /proc/sys/kernel/threads-max
```

or calling the omp library function

```
int omp_get_max_threads()
```

The threads factory

Looking at the output of `00_stack_and_scope` example, you can re-construct how the virtual space of the father process is divided:



The threads factory

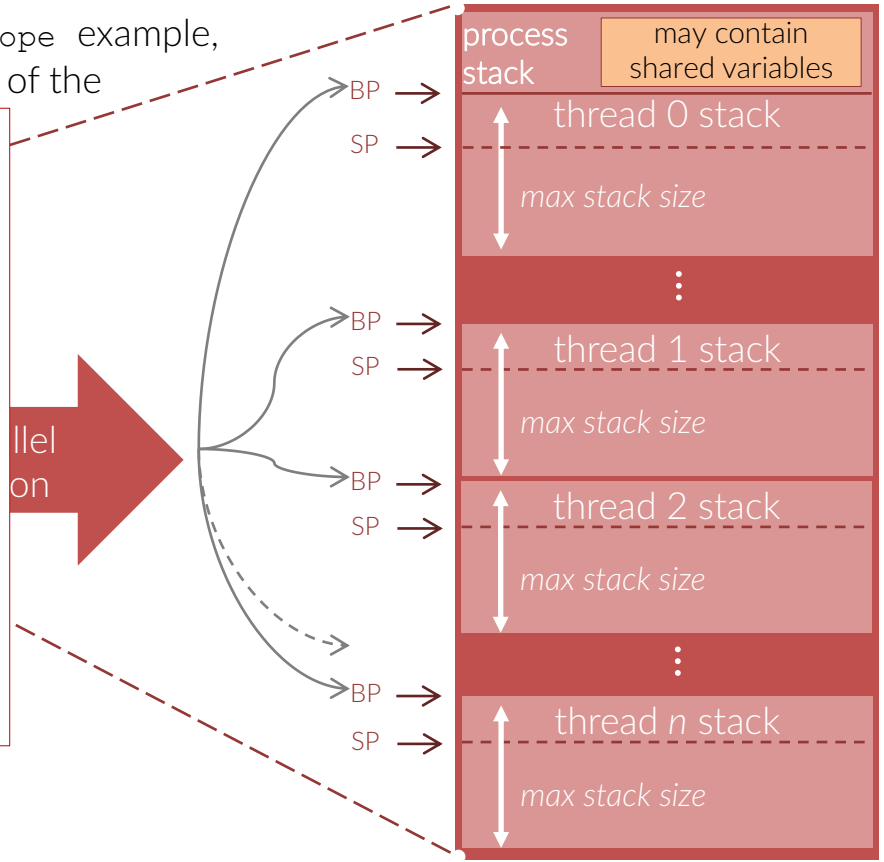
Looking at the output of `00_stack_and_scope` example, you can re-construct how the virtual space of the

Be aware:

When the threads are created, the space needed for their stack is also allocated. If that space is quite large, you may run out of memory.

Conversely, if the stack is not large enough, you may incur into `seg fault` error (but due to lack of memory in the stack)

process' space



Parallel Regions

returns the ID of the calling thread

```
int nthreads     = 1;
int my_thread_id = 0;

#ifdef _OPENMP
#pragma omp parallel
{
    my_thread_id = omp_get_thread_num();
    #pragma master
    nthreads     = omp_get_num_threads();
}
#endif
```

returns the number of threads active in that parallel region

By default rules, variables declared outside a parallel region are *shared*, whereas those declared inside a region are *private*.
Then both `nthreads` and `my_thread_id` are shared.

As a consequence, in this example:
- all threads are writing in `my_thread_id`, in undefined order
- only the master thread is writing in `nthreads`

The value of `my_thread_id` is unpredictable, because it depends on the run-time order by which the threads access it and by each a thread accesses it again to write it.

parallel_regions/
01_simple_pr_wrong.c
02_simple_pr.c

Controlling the number of threads

By default the number of threads spawned in a parallel regions is the number of cores available.
However, you can vary that number in several way

- `OMP_NUM_THREADS environmental variable`
- `#pragma omp parallel num_threads(n)`
- ```
  omp_set_num_threads(n);
  #pragma omp parallel
  ```

`parallel_regions/`
`03_num_of_threads.c`
  ```

Controlling the number of threads

By default the number of threads spawned in a parallel regions is the number of cores available.
However, you can vary that number in several way

- `OMP_NUM_THREADS` environmental variable
- `#pragma omp parallel num_threads(n)`
- `omp_set_num_threads(n);`
  `#pragma omp parallel`

That works if `OMP_DYNAMIC` variable is set to `TRUE`, otherwise the number of threads spawned is strictly equal to `OMP_NUM_THREADS`.
This setting can be changed through **`omp_set_dynamics( true )`**

`parallel_regions/`
`03_num_of_threads.c`

The ordering of the threads


```
#pragma omp parallel
{
    int my_thread_id = omp_get_thread_num();
    printf( "greetings from thread num %d\n",
            my_thread_id );
}
```

The order by which the greetings appear in the output is not deterministic.

In general, unless either you explicitly requires so – and that is possible only for some constructs, or you directly code it, **there is no given order, since the threads are independent entities**.

How would you build-up an enforcement of the order in the parallel region here on the left ?

The ordering of the threads


```
int order = 0;

#pragma omp parallel
{
    int myid = omp_get_thread_num();

    #pragma omp critical
    if ( order == myid ) {
       printf( "greetings from thread num %d\n",
               my_thread_id );
        order++; }
}
```

The CRITICAL directive defines a region which is executed by all threads but by a single thread at a time. Which starts to sound like an “ordering”.

However, although the threads queue at the begin of the region, no particular order is imposed to that queue. As a consequence, the result is unpredictable, aside the fact that `thread 0` will print the message and increase `order`.
For instance, if `thread 2` is queued immediately after, it will execute the `if` evaluation which will fail (`order` would be equal to 1) and the `thread 2` will not print the message.

parallel_regions/
04_order_of_threads_wrong.c

The ordering of the threads


```
int order = 0;

#pragma omp parallel
{
    int myid = omp_get_thread_num();
    int done = 0

    while (!done) {
    #pragma omp critical
    if ( order == myid ) {
       printf( "greetings from thread num %d\n",
               my_thread_id );
        order++; done=1; } }
}
```

parallel_regions/
05_order_of_threads.c

In this second implementation, the `while` cycle enforces the threads that fails the if condition to keep trying until the correct order is ensured.

Once a thread has executed the `critical` region, it then exits the `while` and join the implicit synchronization barrier at the end of the parallel region.

A note: without the `critical` directive, this code could work as well on most platform. However, it is still unreliable since it is not guaranteed that each thread complete the region *before* the successive threads enter in it. For instance, the compiler could have been reshuffled the `order++` to be before the print and so the `thread` *n+1* could enter the region before of the print of the `thread n` and its message could appear as first.

Specializing execution in PR


```
#pragma omp parallel
{
    #pragma omp critical
    { code block
    }

    #pragma omp atomic
    assignment instruction

    #pragma omp single
    { code block
    }

    #pragma omp master
    { code block
    }
}
```

The **`critical`** directive ensures that only a thread at a time executes the block. All the threads will execute it, although in unspecified order.

Like **`critical`** but limited to the following single line.
The instruction can only be an assignment in the form

`x binop = expr`

Only one among the threads will enter the code block

Only the master thread will enter the code block

Specializing execution in PR


```
#pragma omp parallel
{
    #pragma omp critical
    { code block
    }

    #pragma omp atomic
    assignment instruction

    #pragma omp single
    { code block
    }

    #pragma omp master
    { code block
    }
}
```

A barrier (in the form of queuing) is present at the entry point; no one at the end point, i.e. the threads exiting the region will continue the run.

Synchronization as in **`critical`** region.

No explicit synchronization at all (but see relative tips in the following slides).
Only the executing thread is inside the region, the other ones can be everywhere else.

Specializing execution in PR

Rationale of

```
#pragma omp atomic
assignment instruction
```

When you deal with shared variables, ensuring that the workflow maintains the semantic correctness of your code is fundamental.
For instance, the assignment

```
a += b
```

(where either `a`, `b` or both are shared) is meaningful only if the value of `a` (or `b`, or both) does not change in the middle of the instruction itself.

In fact, while a single assembly instruction is guaranteed not to be ever interrupted on the fly, we have to take into account that even a single C line translates in several asm instructions. So, the value of a shared variable could have been changed in the middle of that groups of asm instructions, breaking the correctness of the code.

To avoid that, it is necessary to "protect" the sensible regions.

An **`atomic`** assignment has, obviously, a much lower overhead than a **`critical`** region and must be preferred in the appropriate cases.

Specializing execution in PR

Rationale of

```
#pragma omp atomic
assignment instruction

Clauses:
read, write, update,
capture
```

Special clauses for **`atomic`**

`read`	`var = mem`	causes an atomic read of the location designated by *x* regardless of the native machine word size
`write`	`mem = var`	causes an atomic write of the location designated by *x* regardless of the native machine word size
`update`	`mem++, ++mem, mem--, --mem, mem binop= expr`	Causes an atomic update of the location designated by *x* using the designated operator or intrinsic
`capture`	`val = mem++, val = ++mem, val = mem--, val = --mem, val = mem binop= expr {val = mem; mem binop= expr} {mem binop= expr; val = mem}`	Causes an atomic update of the location designated by *x* using the designated operator or intrinsic while also capturing the original or final value of the location designated by *x* with respect to the atomic update. The original or final value of the location designated by *x* is written in the location designated by *v*, depending on the form of the `atomic` construct, structured block, or statements, following the usual language semantics

Specializing execution in PR


```
#pragma omp critical
{
   "A" code block
}
#pragma omp critical
{
   "B" code block
}
#pragma omp critical(my_loop)
{
   code block
}
```

Named critical regions

In OpenMP it exists a unique global `critical` section.
Hence, when you define a `critical` section, it is logically considered to be part of the global one.
As a consequence *only one thread can be inside any of the unnamed `critical` sections*, which of course limits the perfomance whn more than one region is present.

However, that can be cured by the *named regions*, defined as the last one in the code snippet here on the left.
In that example, only one thread can be either in region "A" or in region "B": so, if one thread is executing "A", another one is forced to delay entering "B" even it would have been reading for it.

Instead, the named `critical` regions can be executed at the same time (by different threads, of course, and only by one thread at a time).

Specializing execution in PR


```
#pragma omp critical
{ …
  some_assignment_to x;
  …
}

#pragma omp atomic
x op= value
```

`critical` *and* `atomic`

Consider that by design `atomic` protects *memory regions*, while `critical` protects code regions.

Hence they are not mutually exclusive: a thread may be executing the critical region while another may be executing the atomic.

Specializing execution in PR


```
#pragma omp single
{ …code block… }
```

```
#pragma omp master
{ …code block… }
```

Is there any difference between using `single` or `master` ?

There obviously is if you're assigning some special workflow to `thread 0`.
If not, the entity of the difference depends on the implementation details.

In general, using `master` requires only a simple test on the thread id, while using `single` requires more synchronization (the `openmp` infrastructure has to keep track about what thread is in the region and so on).

In general, if you have some insight about the fact that the workload before that point may be systematically unbalanced, so that some thread will likely arrive first, using `single` is ok.
Otherwise, it may be better to use `master`.

Work assignment

Inside a parallel region, the work can be assigned to each threads (actually, that's why PR are created..):

```
…
results[0] = heavy_work_0();
results[1] = heavy_work_1();
results[2] = heavy_work_2();
…
```

parallel_region/
06_assign_work.c

```
#pragma omp parallel
{
    if( myid % 3 == 0)
        result = heavy_work_0( );
    else if ( myid % 3 == 1 )
        result = heavy_work_1( );
    else if ( myid % 3 == 2 )
        result = heavy_work_2( );

    if ( myid < 3 )
        results[myid] = result;
}
```

dynamic extent

parallel_region/
06_assign_work.c

run with a second argument >0 to check about it

Conditional creation of PR

It is possible to spawn a parallel region only if some conditions are met:

- `#pragma omp parallel if( any valid C expression )`

```
#pragma omp parallel if( amount_of_work > high )
{
    if( myid % 3 == 0)
        result = heavy_work_0( );
    else if ( myid % 3 == 1 )
        result = heavy_work_1( );
    else if ( myid % 3 == 2 )
        result = heavy_work_2( );

    if ( myid < 3 )
        results[myid] = result;
}
```

If the condition is not fulfilled, the threads are not created and only the master thread will execute the code block.

parallel_region/
06_assign_work.c

if the first argument, that determines the amount of work, is ≤10, the parallel region is **not** created

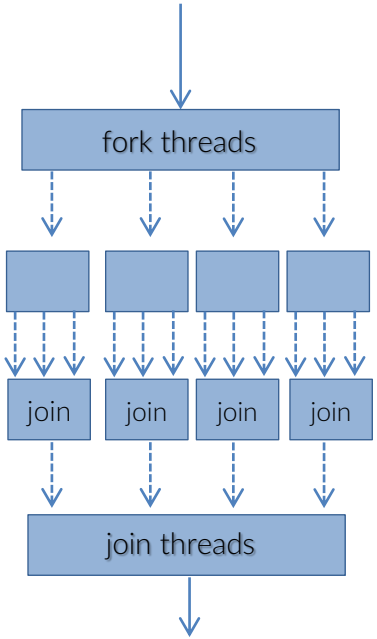
What is this useful for?

The overhead of the creation of a parallel region is of the order of ~10 μseconds (that figure is dependent on the system, the compiler, the number of threads,...).

Then, you may want to create a parallel region (i.e. to spawn more threads) only if the serial execution of that code section is at least several times this overhead.

Nested Parallel Regions

NESTED PARALLELISM



Nested parallelism is explicitly_permitted in OpenMP. Whether it is *active* or not, depends on the value of some environmental variables:

`OMP_NESTED=<TRUE | FALSE>` (*) default value is `FALSE`
`OMP_NUM_THREADS=`N_0`<,`N_1`, … >`
`OMP_MAX_ACTIVE_LEVELS=`*n*

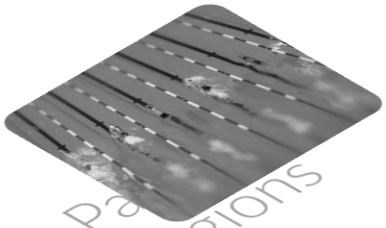
Which you can change by calling the appropriate `OMP_` functions

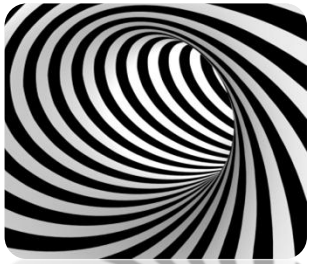
`omp_get_nested(), omp_set_nested(` ***cond*** **`)`**

(*)OMP_NESTED is deprecated in new OpenMP 5.0. You should use only OMP_MAX_ACTIVE_LEVELS and OMP_NUM_THREADS. OMP_NESTED is in fact redundant. However, at the moment of writing the compilers still support OMP_NESTED.

07_nested_regions.c

OpenMP Outline





Parallel Loops

Advanced Parallelism

NUMA AWARENESS

OpenMP parallel loops

Loops are one of the most common work structure in HPC, and it is quite common that a vast amount of compute-intensive code resides in loops.

In fact, OpenMP, up to version 2.x, was essentially about quickly and effectively parallelizing loops without much effort.
Hence, OpenMP standard presents a broad amount of features dedicated to parallel `for` loops.

Building up a parallel loop


```
int N = some_workload;
#pragma omp parallel
{
  int myid = omp_get_thread_num();
  int team = omp_get_num_threads();

  int size     = N / team;
  int reminder = N % team;
  int mystart  = size*myid + (myid<reminder)*myid;
  int myend    = size + (myid < reminder);

  printf("task %d is running from %d to %d\n",
         myid, mystart, myend);

  for ( int i = mystart, i < myend; i++ )
    {
      do_smething(i);
    }
}
```

Splitting the work of a for loop among the threads can easily be achieved by directly assigning the boundaries of the loop to each thread.

In this example, we statically assign an equal share `N/nthreads` of iterations per thread, while distributing the remaining `N%nthreads` iteration to the first `N%nthreads` threads.

However, OpenMP has dedicated constructs that offer easier and more flexible mechanisms to share the work within a for loop.

OpenMP basic loop

Let's start with a classical and very common problem in order to understand the appropriate OpenMP work-sharing construct relative to loops.

```
double *a;
double  sum = 0
int     N;
…
for ( int i = 0; i < N; i++ )
      sum += a[i];
```

OpenMP basic loop


```
#include <omp.h>
double *a, sum = 0;
int     i, N;

#pragma omp parallel for implicit(none) shared(a,sum,N) private(i)
for ( i = 0; i < N; i++ )
      sum += a[i];
```

declares what variables are private: despite their name is the same within the parallel region, they have different memory locations and die with the parallel reg.

This is a **work-sharing** construct; workload is subdivided among threads (the default choice is implementation-dependent)

no implicit assumptions about variables scope

declares what variables are shared; all threads can access and modify those memory locations

OpenMP basic loop

However, variables defined (outside)within the parallel region are automatically (shared) private, and so are the integer indexes used as cycles counter.

```
#include <omp.h>
double *a, sum = 0;
int     N;

#pragma omp parallel for
for ( int i = 0; i < N; i++ )
      #pragma omp atomic
      sum += a[i];
```

parallel_loops/
00_array_sum_with_race.c

What happens if you drop the `atomic` directive?
You obtain a result that is smaller than the correct one: why?

How is the work assigned to single threads ?

OpenMP basic loop


```
#pragma omp parallel for implicit(none) shared(a,sum,N) private(i)
```

The default policy for memory regions is actually that all are shared. However, that is a very common source of error – when you have lots of variables, you forgot what is what in your code.

It may be considered a good practice to add `implicit(none)` to all your construct so that to spot any error alike.

OpenMP basic loop


```
#include <omp.h>
double *a, sum = 0;
int     N;

#pragma omp parallel for
for ( int i = 0; i < N; i++ )
    sum += a[i];
```

Without the `atomic` directive, the assignment

```
sum += a[i];
```

determines a **data race**: between two synchronization points at least one thread writes to a data location from which another threads reads.

Anatomy of a data race

TBD

(check your notes from lecture)

After having solved the data race

Let's say that we solve the data race introducing the critical region `local_sum`, or an `atomic` directive. Does it scale ?

```
#include <omp.h>
double *a, sum = 0;
int     N;

#pragma omp parallel for
for ( int i = 0; i < N; i++ )
      #pragma omp critical local_sum
      sum += a[i];
```

parallel_loops/
00_array_sum_with_race.c

Try to run it with a fixed, large enough, N on an increasing number of cores, and take note about the speedup. Then, measure the Parallel overhead

Of course no! why?

Solving the reduction

Of course no! why?

Because this solution makes the threads to wait for each other too frequently.
A critical region has **synchronization points** at the start and the end of critical regions, meaning that threads have to communicate with each other and decide who's waiting and who's not.

Other **sync points** are implicit and explicit barriers, locks and flush directives.

Solving the reduction / 2

However, that is so important that the OpenMP standard offers a simple solution:

parallel_loops/
01_array_sum.c

```
#include <omp.h>
double *a, sum = 0;
int     N;

#pragma omp parallel for reduction(+: sum)
for ( int i = 0; i < N; i++ )
   sum += a[i];
```

Note that *shared* clause has disappeared; implicit assumptions are ok for us.. in this simple case.

Solving the reduction / 3

There is another way in which we can solve the conflicts on the `sum`

```
#include <omp.h>
double *a;
int     N;
int     nthreads;

#pragma omp master
nthreads = omp_get_num_threads();

double sum[nthreads];

#pragma omp parallel
{
   int me = omp_get_thread_num();
   #pragma omp for
   for ( int i = 0; i < N; i++ )
     sum[me] += a[i];
}
```

Does this scale ?

Solving the reduction /4

There is another way in which we can solve the conflicts on the `sum`

parallel_loops/
02_falsesharing.c

```
#include <omp.h>
double *a;
int     N;
int     nthreads;

#pragma omp master
nthreads = omp_get_num_threads();

double sum[nthreads];
#pragma omp parallel
{
   int me = omp_get_thread_num();
   #pragma omp for
   for ( int i = 0; i < N; i++ )
      sum[me] += a[i];
}
```

Does this scale ?

Hardly

Because the values of **sum[nthreads]** reside in the same cache line(s); hence, when a thread access and modify its location, to maintain the coherence the cache must write-back and reflush. Every time.

That is called **false sharing**

Solving the reduction /5

There is another way in which we can solve the conflicts on the `sum`

parallel_loops/
03_falsesharing_
fixed.c

```
#include <omp.h>
double *a;
int     N;
int     nthreads;

#pragma omp master
nthreads = omp_get_num_threads();

double sum[nthreads*8];
#pragma omp parallel
{
   int me = omp_get_thread_num();
   #pragma omp for
   for ( int i = 0; i < N; i++ )
      sum[me*8] += a[i];
}
```

Does this scale ?

Better.

However, we are using much more memory than needed.
And, above all, we hard-coded a magic number (which is not a good move, in general, since it is not portable).

A subtlety to note

These two "i"s are two different variables, although both are thread-private.

```
#include <omp.h>
#pragma omp parallel
{
    int i;
    #pragma omp for
    for ( i = 0; i < N; i++ )
        {
            …;
        };
}
```

parallel_loops/
01b_array_sum.c

```
luca@GGG:~/work/TEACHING/CODES/OpenMP/parallel_loops$ ./01b_array_sum
omp summation with 4 threads
thread 0 : &i is 0x7ffc27c4b954
        thread 0 : &loopcounter is 0x7ffc27c4b958
thread 1 : &i is 0x7f59f56c3ae4
        thread 1 : &loopcounter is 0x7f59f56c3ae8
thread 2 : &i is 0x7f59f52c1b64
        thread 2 : &loopcounter is 0x7f59f52c1b68
thread 3 : &i is 0x7f59f4ebfb64
        thread 3 : &loopcounter is 0x7f59f4ebfb68
Sum is 4950, process took 0.000830412 of wall-clock time
```

OpenMP work assignment in loops

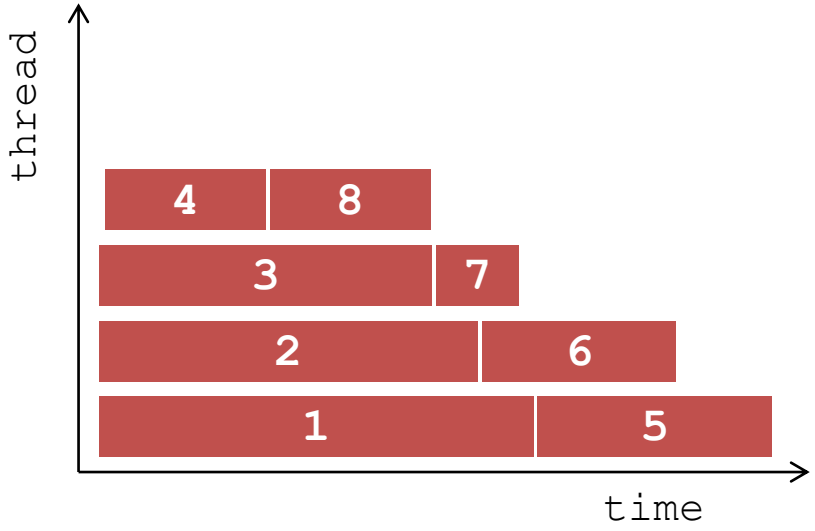
How the work is assigned to the single threads ?

```
#pragma omp parallel for schedule(scheduling-type)
for ( int i = 0; i < N, i++ )
```

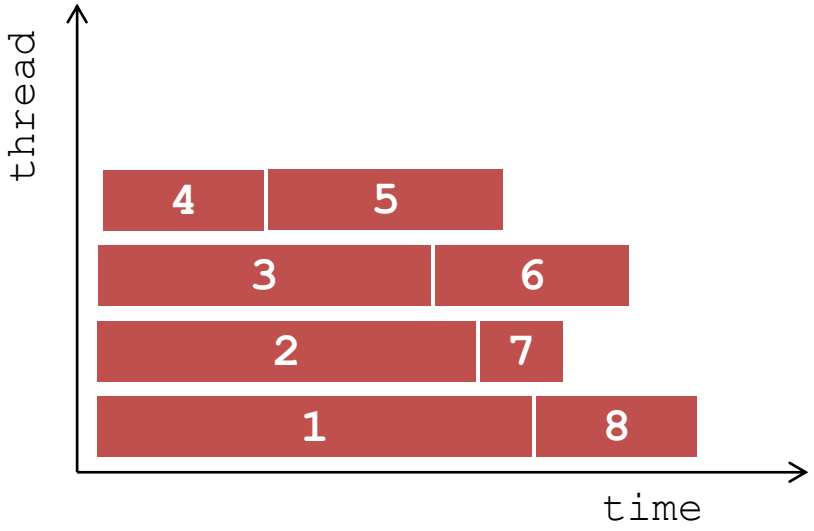
schedule(**static**, *chunk-size*)	The iteration is divided in chunks of size *chunk-size* (or in ~equal size) distributed to threads in circular order
schedule(**dynamic**, *chunk-size*)	The iteration is divided in chunks of size *chunk-size* (or size 1) distributed to threads in no given order (a thread requests the first available chunks)
schedule(**guided**, *chunk-size*)	The iteration is divided in chunks of minimum size *chunk-size* (or size 1) distributed to threads in no given order like *dynamic*. The chunk size is proportional to the number of unassigned iterations divided by the number of threads.
runtime , default	The policy is set at runtime via env. OMP_SCHEDULE or to intern. var. def-sched-var.

OpenMP work assignment in loops

Static vs Dynamic



Static assignment



Dynamic assignment

Clauses in *parallel for*


```
#pragma omp for
        schedule( policy [,chunk])
        ordered
        private ( var list )
        firstprivate ( var list )
        lastprivate (var list )
        shared ( var list )
        reduction ( op: var list )
        collapse (n)
        nowait
```

Clauses in *parallel for*

`private ( var list )`

vars in the list will be private to each thread; despite their name is the same out of the parallel region, they have different memory locations and die with the parallel region.

`firstprivate ( var list )`

the variables in the list are private (in the same sense than in *private*) and are initialized at the value that shared variables have at the begin of the parallel region.

`lastprivate ( var list )`

the shared variables will have the value of the private var in the last thread that ends the work in the parallel region.

Clauses in *parallel for*

firstprivate & lastprivate

```
double PI              = 3.1415blablabla;
int    morning_coffees = MAX_INTEGER;
char   password[]      = “dont_ask_dont_tell”
int    final_mark;

#pragma omp parallel firstprivate( pi, morning_coffees) private(password) lastprivate( final_mark)
{
    drink_mycoffees( morning_coffees );
    use_pi( PI );

    password = setup_mypassword();
    int exam_passed = 0;
    while (!exam_passed) { exam_passed = try() }

    final_mark = exam_passed;
}
```

Clauses in *parallel for*

`reduction ( op: var list )`

Possible operators are: +, ×, -, max, min, &, &&, |, ||
The initial value of vars is taken into account *at the end* of the parallel for; at the begin of the for, initialization values are what you logically expect: 0 for add, 1 for mul, min and max of the result type for max and min.

`collapse ( n )`

Enable the parallelization of multiple loops level (must be perfectly nested)

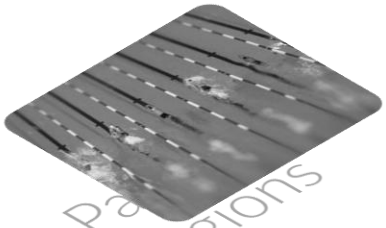
```
#pragma omp for collapse(2)
for ( int ii = 0; ii < Nrows; ii++ )
   for ( int jj = 0; jj < Ncol; jj++ )
      A[i][j] = B[i][j] * C[i][j];
```

```
#pragma omp for collapse(2)
for ( int ii = 0; ii < Nrows; ii++ ) {
   D[i] = function_of_(i);
   for ( int jj = 0; jj < Ncol; jj++ )
      A[i][j] = B[i][j] * C[i][j] + D[i]; }
```

`nowait`

Ignore the implicit barrier at the end of parallel region or work-sharing construct

OpenMP Outline



Advanced Parallelism

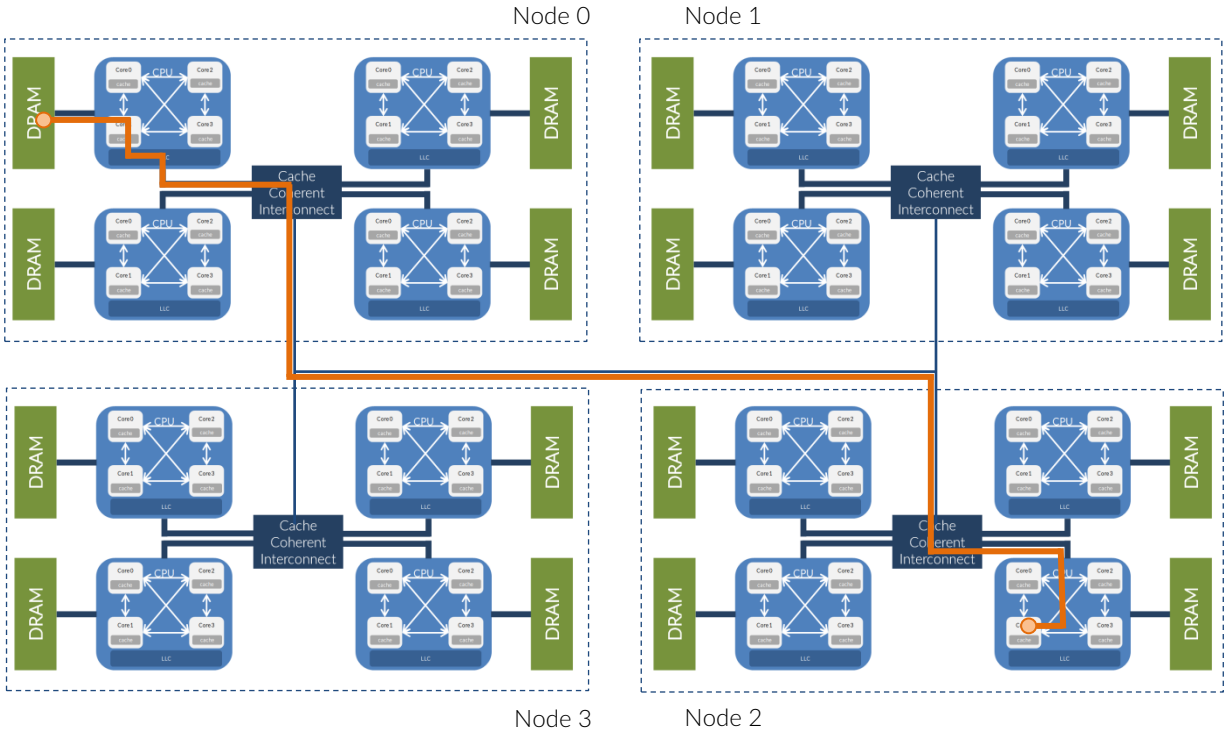
NUMA AWARENESS

NUMA Outline

- The problem: “Where? Who? What?”
- Touch-first and touch-by-all policy
- Threads affinity

Remind: get back to The typical NUMA architecture

Where? Who? What?



The aim is to have as few *remote memory accesses* as possible. That depends on

- **Where**: i.e. in what memory bank the data are;
- **Who**: is accessing them, i.e. which thread → how are the threads distributed on the cores;
- **What**: how is the work-load distributed among the threads;

Where do the threads run ?

OpenMP and the OS offer the capability to decide where each thread have to run, i.e. on which core and/or how the threads have to distribute on the available cores.

We know that each core may have the capability of running more than one thread, which is called (*) Simultaneous MultiThreading (SMT).
In the next slides, let's call *strands* or *hwthreads (hardware threads)* the different threads that a physical core could run, as opposed to "*swthreads*" (software threads) that indicates the OpenMP threads.

The placement of OpenMP threads on cores is called "**threads affinity**".

(*)The *Hyper-Threading* is the proprietary technlogy, and trademark, of Intel's SMT

In principle, you want to be able to distribute the work in an optimal way, i.e. without any resource (computational power, caches and memory) contention.

To do that, you must be able to place each OpenMP *swthread* to a dedicated computational resource, and to grant it the fastest possible access to "its own" data.

So, you need to:

- explicitly bind the threads to "cores", i.e. *hwthreads*
- explicitly allocate memory on the best suited physical memory
- minimize the remote memory access
- In case, to migrate memory and/or *swthreads* to one NUMA node to another, or to one *hwthreads* to another respectively

OpenMP offers 2 basic concepts to set and control the affinity:

PLACES

i.e. to what physical entities (*hwthread*) we are referring to with our affinity request: "where" the threads run.

BINDING

i.e. whether there is some request about the relationship between threads and PLACES (in other words: between swthreads and hwthreads), and what that request is.

Threads affinity - RATIONALE

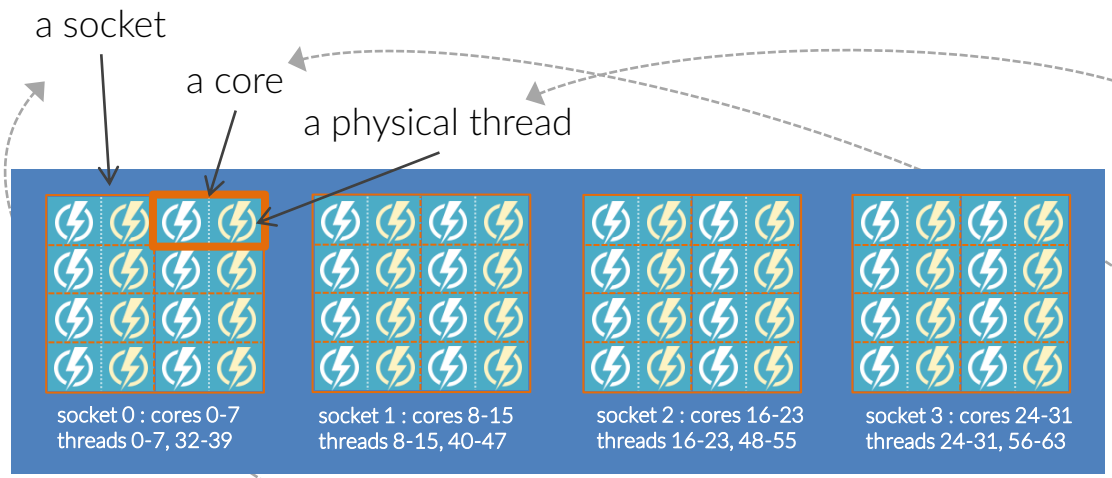
What is the "optimal" way to place the swthreads in a node depends on the nature of the algorithm and the data you are dealing with.

Having the swthreads "**distant**" from each other:

- may increase the aggregate bandwidth – i.e. each hwthread could fully exploit its available bandwidth – *if* the data are placed accordingly;
- may result in a better utilization of each core's cache, because it would be reserved to a single swthread's data;
- may worsen the performance of synchronization constructs.

Symmetrically, having the swthreads "**close**" to each other:

- may decrease the latency of synchronization constructs;
- may decrease the aggregated bandwidth;
- may worsen the cache performance.



A typical example of configuration for a multicore, multisocket node: 4 sockets, each with 8 cores, each with 2 hwthreads.
Physical threads are exposed as “cores”, numbered in a round robin fashion.

PLACES are where swthreads run.
The names for PLACES are:

- THREADS each place corresponds to a hwthread, or strand, on cores
- CORES each place corresponds to a single core (which may have more strands) on sockets
- SOCKETS each place corresponds to a physical sockets, with its multiple cores

A “place” can be defined by an **unordered** set of comma-separated non-negative numbers enclosed in braces (*the numbers are the IDs of the smallest unit of execution on that hardware, usually a hwthread*).
Comma separated places define an ***ordered*** *list*.

{ 0, 1 }	this defines a place made by hwt 0 and hwt 1 *in the frame of the previous examples, these are the hwt on core 0 and core 1 of socket 0*
{ 0, 49 }	this defines a place made by hwt 0 and hwt 49 *in the frame of the previous examples, these are the hwt and the SMT hwt on core 0 of socket 0*
{ 0, 12, 24, 36 }	this defines a place made by hwt 0 and hwt 1 *in the frame of the previous examples, these are the hwt on cores 0 of sockets 0, 1, 2 and 3*
{ 0,1 }, { 0 ,49 }	A list with two places

OMP_PLACES can be defined as an explicit ordered list of comma-separated places (see the previous slide for a definition of “places”).

Intervals can de used, specified as `start:counter:stride` which results in the serie

```
start, start+stride, start + 2×stride, …, start + (counter-1)×stride
```

OMP_PLACES = { 0, 48} , {1, 49}	sets OMP_PLACES to 2 places *in the frame of the previous examples, these are the hwt and SMT hwt on cores 0 and 1, respectively, of socket 0*
OMP_PLACES = { 0:2:48}, {1:2:49}	the same than previous line
OMP_PLACES = { 0, 12, 24, 36 }	SET OMP_PLACES to 1 place *in the frame of the previous examples, these are the hwt on cores 0 of sockets 0, 1, 2 and 3*
OMP_PLACES = { 0:4:12 }	the same then previous line

Other examples of places definition by intervals:

{ 0 }:4:12	{ 0 }, { 12 }, { 24 }, { 36 }
{ 0:4:1 }:4:12	{ 0,1,2,3 }, { 12,13,14,15 }, { 24,25,26,27 }, { 36,37,38,39 }
{ 0:4}:4:4 { 0:4 }, { 4:4 }, { 8:4 }, { 12:4 }	{ 0,1,2,3 }, { 4,5,6,7 }, { 8,9,10,11 }, { 12,13,14,15 }
{ 0:12 }:4:12	Equivalent to OMP_PLACES=`sockets` on a system with 4 sockets with 12 cores each

The ! Operator can be used to *exclude* intervals.

The places are *static*: there is no way to change it while the program is running.

It some of the specified places is not available, the behaviour is implementation dependent.

Threads affinity - BINDING

The BINDING defines how the swthreads are mapped onto the PLACES.

The names for BINDING are:

- NONE the placement is up to the OS
- CLOSE the swthreads are placed onto places as close as possible to each other
- SPREAD the swthreads are placed onto places as far as possible to each other, then filled in a round-robin fashion
- MASTER the swthreads run onto the same place than master thread

The binding can be specified in a non-persistent way for each parallel region *inside* the code:

```
#pragma omp parallel proc_bind(policy)
```

Once a swthread has been assigned to a hwthread, it is not allowed to migrate. If you have *nested parallelism*, you may define different behaviour for the nested regions

```
#pragma omp parallel proc_bind(spread)
{
   #pragma omp parallel for proc_bind(close)
   for ( int ii = 0; ii < local_N; ii++ )
}
```

Threads affinity - BINDING


```
export OMP_PLACES=threads
export OMP_NUM_THREADS=4,10
export OMP_PLACES=spread,close
```

close

- T ≤ P : there are sufficient places for a unique assignment.
  Swthreads are assigned to *consecutive places* by their thread ID.
  The first place is the master's place.
- T > P : at least one place executes more than one swthread.
  Swthread are splitted in P subsets St_i, so that
  $floor(T/P) \leq St_i \leq ceiling(T/P)$
  St_0 includes swt 0 and is assigned to master's place.

- T ≤ P : place list is splitted in subpartitions; each subpartition contains at least *floor(P/T)* and at most *ceiling(P/T) consecutive places*. A thread is then assigned to a subpartition, starting from the master thread. Then, assignment proceeds by thread ID, and the threads are placed in the first place of the next subpartition.

spread

- T > P : place list is splitted in subpartitions, each of which contains only 1 place and St_i *threads* with consecutive IDs.
  The number of threads St_i in each subpartition is chosen so that:
  $floor(T/P) \leq St_i \leq ceiling(T/P)$
  At least one place has more than one thread assigned to it.
  The first subset with St_0 contains thread 0 and runs on the place that hosts the master thread.

Threads affinity - examples

binding \ places	THREADS	CORES	SOCKETS
CLOSE	swt are placed on close hwt, saturating all the SMT hwt in each core before using new cores	swt are placed on close hwt, using 1 hwt/core before starting to use SMT	swt are placed round-robin per socket, 1/core; after saturation, SMT is used by round-robin +1 hwt/socket
SPREAD	swt are placed round-robin sockets, onto free cores in sockets	similar to ← SMT is avoided until saturation	similar to ← swt are placed by round-robin sockets and hwt
MASTER	all swt are placed on the same hwt on the same core on the same socket	all swt are placed on the same core on the same socket, using all its hwt	all swt are placed on the same socket, saturing all hwt starting from SMT ones

note:

```
swt = software threads
hwt = hardware threads
```

Threads affinity - examples


```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                96
On-line CPU(s) list:   0-95
Thread(s) per core:    2
Core(s) per socket:    12
Socket(s):             4
NUMA node(s):          4
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz
Stepping:              4
CPU MHz:               1000.073
CPU max MHz:           3200.0000
CPU min MHz:           1000.0000
BogoMIPS:              4600.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              16896K
NUMA node0 CPU(s):     0-11,48-59
NUMA node1 CPU(s):     12-23,60-71
NUMA node2 CPU(s):     24-35,72-83
NUMA node3 CPU(s):     36-47,84-95
```

First hardware threads on sockets.
Do exist also when SMT is switched off

Second hardware threads.
Depend on SMT

Socket 0 ---->

Socket 1

Socket 2

Socket 3

0	48	4	52	8	56
1	49	5	53	9	57
2	50	6	54	10	58
3	51	7	55	11	59

The following examples will refer to a node like the one reported here on the left:

4 sockets
12 cores / socket
2 hwthreads / core

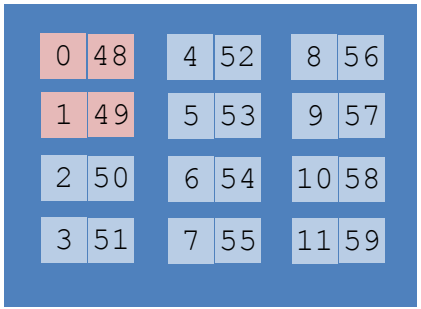
OMP_PLACES = threads

OMP_PROC_BIND = close

swt are placed on close hwt, saturating all the siblings SMT hwt in each core before using new cores

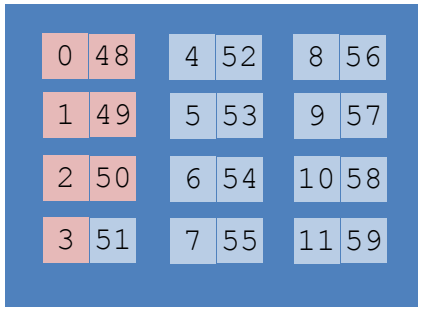
numa_awareness/
00_where_I_am.c

4 swthreads



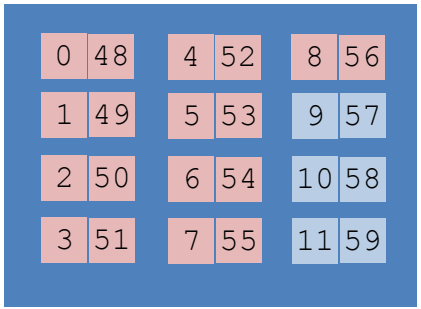
Socket 0

7 swthreads



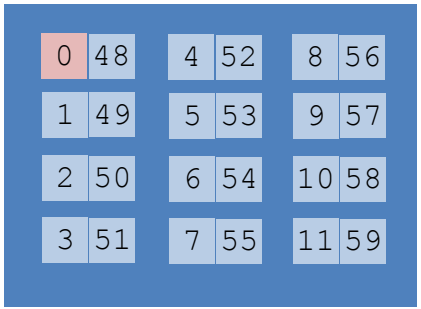
Socket 0

18 swthreads



Socket 0

25 swthreads



Socket 1 (S0 saturated)

Threads affinity - examples

OMP_PLACES = threads

OMP_PROC_BIND = close

swt are placed on close hwt, saturating all the siblings SMT hwt in each core before using new cores

4 swthreads

	hwthreads	swthreads
node 0	00-11 , 48-59	0,48,1,49 ●●○○
node 1	12-23 , 60-71	
node 2	24-35 , 72-83	
node 3	36-47 , 84-95	

swthreads places are reported by ID order.

- ● *means hwthread*
- ○ *means SMT hwthread*

7 swthreads

	hwthreads	swthreads
node 0	00-11 , 48-59	0,48,1,49,2,50,3 ●●●●○○○
node 1	12-23 , 60-71	
node 2	24-35 , 72-83	
node 3	36-47 , 84-95	

18 swthreads

	hwthreads	swthreads
node 0	00-11 , 48-59	0,48,1,49,2,50, 3,51,4,52,5,53, 6,54,7,55,8,56 ●●●●●●●● ○○○○○○○○
node 1	12-23 , 60-71	
node 2	24-35 , 72-83	
node 3	36-47 , 84-95	

numa_awareness/ 00_where_I_am.c

OMP_PLACES = threads

OMP_PROC_BIND = close

swt are placed on close hwt, saturating all the siblings SMT hwt in each core before using new cores

25 swthreads

	hwthreads	swthreads
node 0	00-11 , 48-59	SATURATED
node 1	12-23 , 60-71	12 ●
node 2	24-35 , 72-83	
node 3	36-47 , 84-95	

50 swthreads

	hwthreads	swthreads
node 0	00-11 , 48-59	SATURATED
node 1	12-23 , 60-71	SATURATED
node 2	24-35 , 72-83	24, 72 ●○
node 3	36-47 , 84-95	

numa_awareness/
00_where_I_am.c

Threads affinity - examples

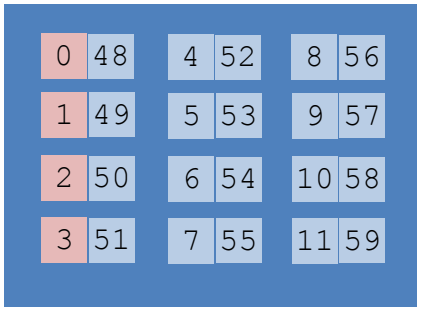
OMP_PLACES = cores

OMP_PROC_BIND = close

swt are placed on close hwt, using 1 hwt/core before starting to use SMT

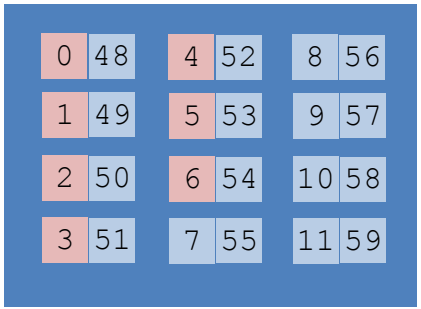
numa_awareness/
00_where_I_am.c

4 swthreads



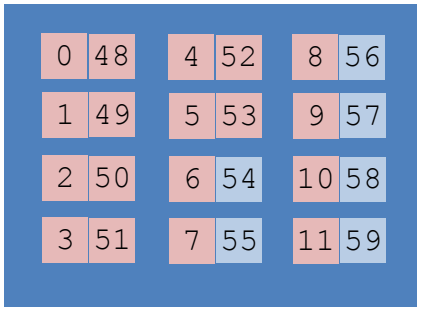
Socket 0

7 swthreads



Socket 0

18 swthreads



Socket 0

OMP_PLACES = sockets
OMP_PROC_BIND = close

swt are placed round-robin per socket, 1/core; after saturation, SMT is used by round-robin +1 hwt/socket

numa_awareness/
00_where_I_am.c

4 swthreads

Socket 0						Socket 1					
0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 2						Socket 3					
24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

OMP_PLACES = sockets

OMP_PROC_BIND = close

swt are placed round-robin per socket, 1/core; after saturation, SMT is used by round-robin +1 hwt/socket

numa_awareness/
00_where_I_am.c

14 swthreads

0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 0 Socket 1

24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

Socket 2 Socket 3

OMP_PLACES = threads

OMP_PROC_BIND = spread

swt are placed round-robin sockets, onto free cores in sockets

numa_awareness/
00_where_I_am.c

4 swthreads

Socket 0		Socket 1		Socket 2		Socket 3	
0	48	12	60	24	72	36	84
1	49	13	61	25	73	37	85
2	50	14	62	26	74	38	86
3	51	15	63	27	75	39	87
4	52	16	64	28	76	40	88
5	53	17	65	29	77	41	89
6	54	18	66	30	78	42	90
7	55	19	67	31	79	43	91
8	56	20	68	32	80	44	92
9	57	21	69	33	81	45	93
10	58	22	70	34	82	46	94
11	59	23	71	35	83	47	95

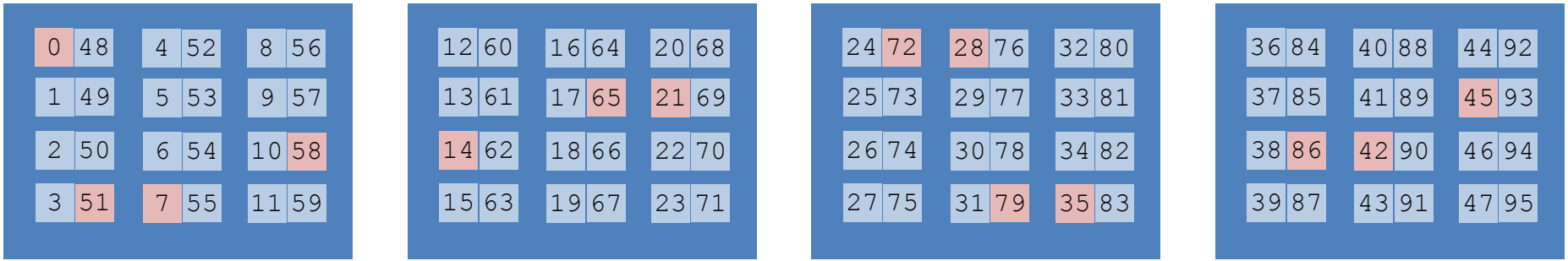
OMP_PLACES = threads

OMP_PROC_BIND = spread

swt are placed round-robin sockets, onto free cores in sockets

numa_awareness/
00_where_I_am.c

14 swthreads



Socket 0 Socket 1 Socket 2 Socket 3

Threads affinity - examples

OMP_PLACES = cores

OMP_PROC_BIND = spread

OMP_PLACES = sockets

OMP_PROC_BIND = spread

Similar to (thread, spread), just infer the differences from the table description. And, run on your own `00_whare_I_am` to check what is happening.

`numa_awareness/`
`00_where_I_am.c`

OMP_PLACES = threads

OMP_PROC_BIND = master

4 swthreads

4 swthreads are running on this same hwthread

Socket 0						Socket 1					
0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 2						Socket 3					
24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

OMP_PLACES = cores

OMP_PROC_BIND = master

4 swthreads

2 swthreads are running on each of these 2 hwthreads

Socket 0						Socket 1					
0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 2						Socket 3					
24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

OMP_PLACES = sockets

OMP_PROC_BIND = master

4 swthreads

0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 0 Socket 1

24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

Socket 2 Socket 3

OMP_PLACES = sockets

OMP_PROC_BIND = master

20 swthreads

0	48	4	52	8	56	12	60	16	64	20	68
1	49	5	53	9	57	13	61	17	65	21	69
2	50	6	54	10	58	14	62	18	66	22	70
3	51	7	55	11	59	15	63	19	67	23	71

Socket 0 Socket 1

24	72	28	76	32	80	36	84	40	88	44	92
25	73	29	77	33	81	37	85	41	89	45	93
26	74	30	78	34	82	38	86	42	90	46	94
27	75	31	79	35	83	39	87	43	91	47	95

Socket 2 Socket 3

OMP_PLACES = sockets

OMP_PROC_BIND = master

35 swthreads

When all the hwthreads are saturated, more than 1 swthread is placed on hwthreads by round-robin, on the same socket

0 48	**4** 52	**8** 56
1 **49**	**5** 53	9 **57**
2 50	**6** 54	**10** 58
3 **51**	7 **55**	11 **59**

Socket 0

12 60	16 64	20 68
13 61	17 65	21 69
14 62	18 66	22 70
15 63	19 67	23 71

Socket 1

24 72	28 76	32 80
25 73	29 77	33 81
26 74	30 78	34 82
27 75	31 79	35 83

Socket 2

36 84	40 88	44 92
37 85	41 89	45 93
38 86	42 90	46 94
39 87	43 91	47 95

Socket 3

Threads affinity

If you compile `00_whare_I_am.c` with `-DSPY` , it will load the hwthreads with some amount of work so that meanwhile you can inspect what is happening by using the top utility:

On my laptop using 2 swthreads

```
top - 14:11:56 up  5:29,  4 users,  load average: 3.01, 2.10, 1.81
Threads: 899 total,   3 running, 828 sleeping,   0 stopped,   0 zombie
%Cpu0  : 16.4 us,  7.0 sy,  0.0 ni, 76.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu1  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu2  :100.0 us,  0.0 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu3  :  9.8 us,  3.9 sy,  0.0 ni, 86.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 27.7/16241208 [
KiB Swap:  0.0/35639292 [

  PID  PPID   UID USER     RUSER    TTY          TIME+ %CPU %MEM S COMMAND
20037 15584  1000 luca     luca     pts/1      1:17.54 99.9  0.0 R 00_where_I_am_s
20036 15584  1000 luca     luca     pts/1      1:14.71 99.9  0.0 R 00_where_I_am_s
 3271  3240  1000 luca     luca     ?          5:24.05  5.3  0.7 S kwin_x11
 2660  2656     0 root     root     tty1      11:13.78  4.3  1.0 S Xorg
 4181  3223  1000 luca     luca     ?          4:05.04  3.9  1.8 S Wavebox
 3279     1  1000 luca     luca     ?          3:20.64  3.0  3.0 S plasmashell
```

Threads affinity – omp functions

The OpenMP standard offers several `omp_` functions to deal with the affinity…

TBD

Memory allocation

1. By carefully touching data
2. By changing default memory allocation with `numactl`
3. By explicit memory migration

… TBCompleted

1. Careful data touching

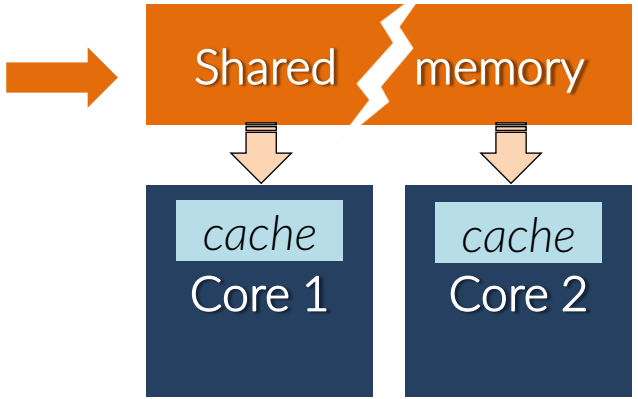
“touch-first” policy

The matter is: who “owns” the data?

```
double *a = (double*)calloc( N, sizeof(double);

for ( int i = 0; ii < N; ii++ ) {
   a[i] = initialize(i);

#pragma omp parallel for reduction(+: sum)
for ( int i = 0; i < N; i++ )
      sum += a[i];
```



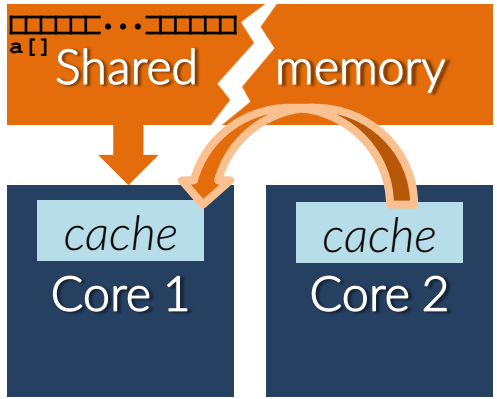
The matter is: who “owns” the data?

parallel_loops/
01_array_sum.c

```
double *a = (double*)calloc( N, sizeof(double);

for ( int i = 0; ii < N; ii++ ) {
   a[i] = initialize(i);

#pragma omp parallel for reduction(+: sum)
for ( int i = 0; i < N; i++ )
      sum += a[i];
```



In this way, the cache of the thread that initialize (first touch) the data is warmed-up **and the data are allocated in the memory connected to it.**

In the “touch-first” policy, the data pages are allocated in the physical memory that is the closest to the physical core which is running the thread that access the data first.
If a single thread is initializing all the data, then all the data will reside in its memory and the number of remote accesses will be maximized.

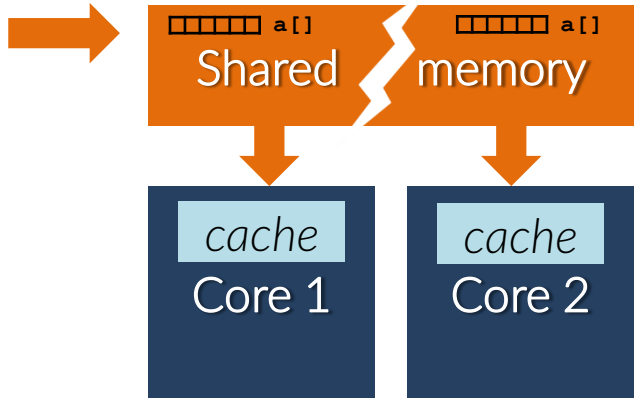
The matter is: who “owns” the data?

parallel_loops/
06_touch_by_all.c

```
double *a = (double*)calloc( N, sizeof(double);

#pragma omp parallel for
for ( int i = 0; ii < N; ii++ ) {
   a[i] = initialize(i);

#pragma omp parallel for reduction(+: sum)
for ( int i = 0; i < N; i++ )
      sum += a[i];
```



In this way, the cache of each thread is warmed-up with the data it will use afterwards **and the data are allocated into each thread’s memory** (the scheduling must be the same!)

“touch-by-all” policy

If each thread “touches” as first the data it will operate on subsequently, those data – by the “touch-first” policy – are allocated in the physical memory that is the closest.
Hence, each thread will have its data placed in the most convenient memory and the remote accesses will be minimized

```
double *global_pointer;
double  global_damnimportant_variable;

#pragma omp threadprivate(global_pointer, global_damnimportant_variable)
```

`threadprivate` preserves the global scope of the variable, but make it private for every thread.
Basically, every thread has its own copy if it everywhere you create a thread team.

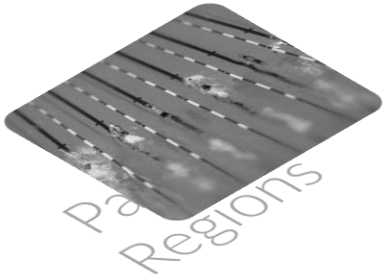
2. Change allocation policy

TBD

3. Explicit memory migration

TBD

OpenMP Outline



Advanced Parallelism

- A parallel construct like `for` amounts to create a “Single Program Multiple Data” instance: all the threads execute the same code but on different data.
- Other work-sharing constructs are instead about assigning different execution paths through the code among the threads.
  - **`section`** construct
  - **`tasks`** construct

A general view (recap)

Multi-cores architectures

Around mid of 2000's, it became clear that speedup applications relying on the scaling-up of CPU's frequency was no longer possible.

- Heterogeneous computing
- Parallel computing — not a direct consequence of the end of "free lunch", but deeply affected by it
- New paradigm for programming
- Increasing fine-grain parallelism

The challenge in writing complex scientific and data-intensive applications has increasingly become manyfold.

(1) to identify the parts of the works that can be parallelized, and to expose that parallelism;

(2) to add additional considerations about resource contention, particularly to concurrent data access

(3) to identify a finer-grained parallelism, decomposing the workflow in smaller well-defined "sequences" of operations that use a subset of data, with well-defined dependencies with other "sequences"

Different approaches

I	Providing concurrency mechanisms through APIs	The programmer is in charge to find and implement the parallelism, and to manage the concurrency	`MPI, POSIX threads early OpenMP`
II	OOO-features and general parallel patterns	The aim is to alleviate the programmer's burden making the technical details as invisible (transparent) as possible	`Intel TBB, HPX, FastFlow, …`
III	Inherent parallel constructs	Native parallelism in the language, improve readability and compactness, support sync and concurrency control. Mostly based on Partitioned Global Address Space (PGAS), which focus on data instead of communications/concurrency controls	`OpenMP, UPC, Chapel, CoArray Fortran,…`
IV	Task-based approach, automatic extraction of parallelism	"tasks" are defined in different ways, and a graph of dependencies is derived (implicitly or explicitly): nodes are the procedures and edges are the relations among them. The "assignment" of data regions to the tasks determines the safe concurrent access.	`Intel TBB, Cilk, Charm++, TensorFlow, StarPU, OmpSs, late OpenMP,…`

Task abstraction

represents any contained sequence of instructions in the code, logically defining a finite work/function/assignment

Asynchronous + Interleaved execution + dependencies

Dependency graph among task.
Must be acyclic.
Related to data dependencies.

Data abstraction

represents any piece of logically uniform “information”, that may be accessed by several threads; out-of-order access needs to be managed

Concurrent access

The easiest way to get MPMD, i.e. different threads executing different pieces of codes (..“sections”) to accomplish different jobs.
It is useful when there is an *established number* of *independent code units* at compile-time:

```
#pragma omp sections clauses…
{
   #pragma omp section
   { code block }

   #pragma omp section
   { code block }
}
```

Independent structured block of codes

Each block is executed by one, and only one thread; the assignment is implementation dependent

If there are *more sections than threads*, some thread executes more than one section
If there are *more threads than sections*, some threads wait at the implied barrier at the end of the sections construct.

The clauses supported by the `sections` construct are the following:

`private`*(list)*
`firstprivate`*(list)*
`lastprivate`*(list)*
`reduction`*(operator:list)*
`nowait`

```
#pragma omp sections clauses…
{
   #pragma omp section
   function_a(…);

   #pragma omp section
   { do_something here
     function_b();
   }

   #pragma omp section
   { do_something here
     function_c();
   }
}
```

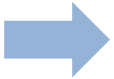
Typical usage of sections is when you have a pre-defined amount of work that can be split in smaller pieces or independent components.

For instance, that may be the case when you have parallel I/O.
Or when you have a computation made up by more independent trunks.

However, the “mechanism” is somehow rigid and may result in a severe work unbalance

OpenMP sections - ex.1


```
for( int ii = 0; ii < N; ii++ )
  result += heavy_work_0(array[ii]) +
    heavy_work_1(array[ii]) +
    heavy_work_2(array[ii]) ;
```



```
#pragma omp sections reduction(+:result)
{

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_0( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_1( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_2( array[jj] );
    result += myresult;
  }

}
```

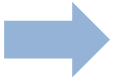
The calls to `heavy_work_?()` are mutually independent. Then, it is possible to separately call the 3 functions for each array entry using the section `construct`.

However, it has no flexibility: the speedup just does not increase while a larger number of threads is used.

parallel_tasks/
00_sections.c

OpenMP sections - ex.1


```
for( int ii = 0; ii < N; ii++ )
  result += heavy_work_0(array[ii]) +
    heavy_work_1(array[ii]) +
    heavy_work_2(array[ii]) ;
```



```
-------------------------------------
executing  00_sections
-------------------------------------

there are 4 NUMA nodes

        running the serial version
                running 0/3
                running 1/3
                running 2/3
34.0559  +-  0.0323754
        running with 3 threads
                running 0/3
                running 1/3
                running 2/3
15.1382  +-  0.0391657
        running with 4 threads
                running 0/3
                running 1/3
                running 2/3
15.0902  +-  0.00148997
        running with 8 threads
                running 0/3
                running 1/3
                running 2/3
15.2701  +-  0.257149
        running with 12 threads
                running 0/3
                running 1/3
                running 2/3
15.3972  +-  0.206368
        running with 16 threads
                running 0/3
                running 1/3
                running 2/3
```

The calls to `heavy_work_?()` are mutually independent. Then, it is possible to separately call the 3 functions for each array entry using the section `construct`.

However, it has no flexibility: the speedup just does not increase while a larger number of threads is used.

parallel_tasks/
00_sections.c

OpenMP sections – ex.2


```
#pragma omp sections reduction(+:result)
{

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_0( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_1( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_2( array[jj] );
    result += myresult;
  }

}
```

We can add some flexibility by spawning a nested parallel region in each section

parallel_tasks/01_sections_nested.c

```
#pragma omp sections reduction(+:result)
{
  #pragma omp  section
  {
    double myresult = 0;
    #pragma omp parallel for reduction(+:myresult)
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_0( array[jj] );
    result += myresult;
  }
  #pragma omp  section
  {
    double myresult = 0;
    #pragma omp parallel for reduction(+:myresult)
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_1( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    #pragma omp parallel for reduction(+:myresult)
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_2( array[jj] );
    result += myresult;
  }
}
```

```
#pragma omp sections reduction(+:result)
{

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_0( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_1( array[jj] );
    result += myresult;
  }

  #pragma omp  section
  {
    double myresult = 0;
    for( int jj = 0; jj < N; jj++ )
      myresult += heavy_work_2( array[jj] );
    result += myresult;
  }

}
```

We can add some flexibility by spawning a nested parallel region in each section

parallel_tasks/
01_sections_nested.c

```
# ------------------------------------
# idx 1
# 01 20k 2k
#   nested parallelism
#   first level is always 3 threads
#

#th     time(s)         std.dev.(s)
1       34.1            0.11
3       15.1            0.05
6       7.98            0.09
12      4.11            0.07
24      2.17            0.03
30      1.79            0.02
```

OpenMP sections


```
int semaphores[3] = {0};
#pragma omp sections clauses…
{
   #pragma omp section
   { while( there_is_work0 ) {
       function_0(…);
       unlock_semaphore(0, …); } }

   #pragma omp section
   { while( there_is_work1 ) {
       check_semaphore(0, …);
       function_1();
       unlock_semaphore(1, …);} }

   #pragma omp section
   { while( there_is_work2 ) {
       check_semaphore(1, …);
       function_2(); } }
}
```

A dependence can be enforced among sections in an explicit way.

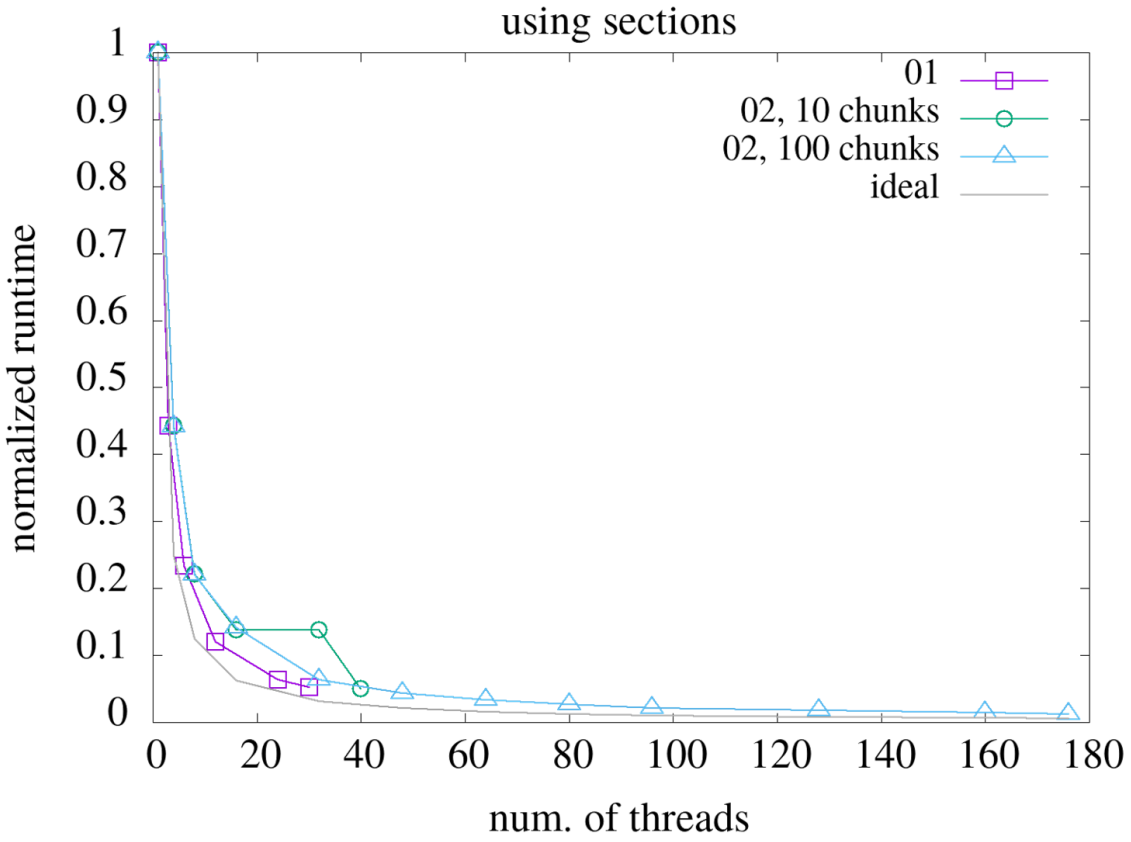
```
void unlock_semaphore( int i, int val ) {
   // in principle you don't need atomic here
   // because each entry should be modified by
   // only one thread
   semaphore[i] = val; }

void check_semaphore( int i, int val ) {
   while( semaphore[i] != val) {
      sleep_or_spin_a_while(…); } }

void sleep_or_spin_a_while( int a_while ) {
   // you may use a call to nanosleep() instead
   volatile int sum = 0;
   for( int ii = 0; ii < N; ii++ ) sum += ii; }
```

OpenMP sections - ex.3

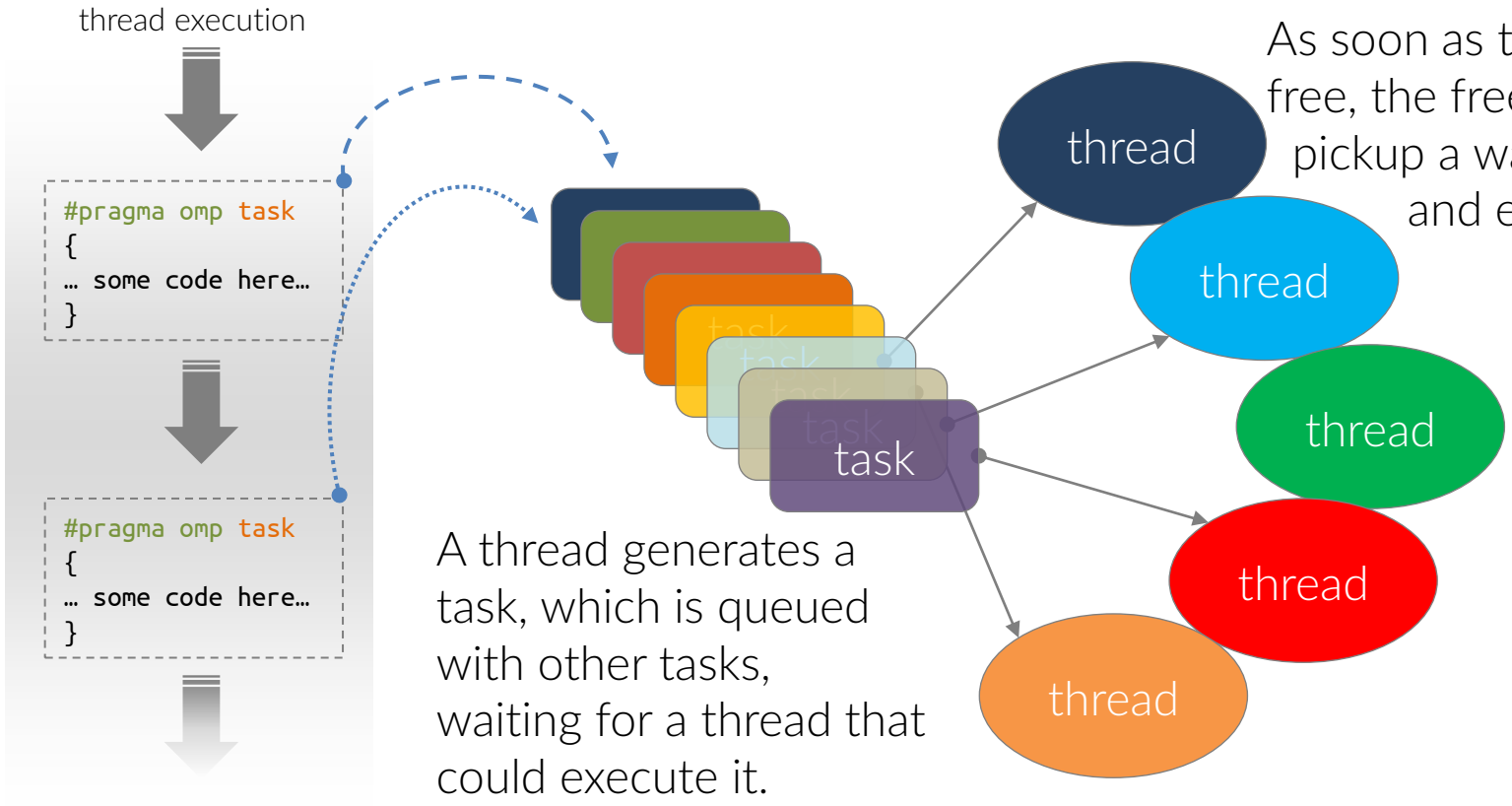
OpenMP sections



As we have seen in the previous example (`02_sections_nested_irrecular.c`), it is sometime possible to parallelize a workflow which is irregular or runtime-dependent.
However, often the solution is quite ugly and convoluted.

Since version 3.0, OpenMP offers a new elegant construct designed for this class of problems.

What happens under the hood is that OpenMP creates a “bunch of work” along with the data and the local variables it needs, and *schedule* it for execution at some point in the future.


thread execution
#pragma omp task
{
… some code here…
}
#pragma omp task
{
… some code here…
}
task
thread
thread
thread
thread
thread
As soon as they are free, the free tasks pickup a waiting task and execute it
A thread generates a task, which is queued with other tasks, waiting for a thread that could execute it.

As almost everything else, a task must be generated *inside* a parallel region and it is linked to a specific block of code.
If its execution is not properly “protected”, it might be executed by *more* than one thread which is not in general what we want.

To guarantee that each task is executed only once, every task must be generated within a `single` or `master` region.

The single region is preferable because of its implied barrier that makes all tasks to be completed before passing. In case you use a `master` region, pay attention to the execution flow.
However, the `master` has often the heavier burden so it’s best to user a `single` region, possibily with the `nowait` clause.

A classical example: traversing a linked list

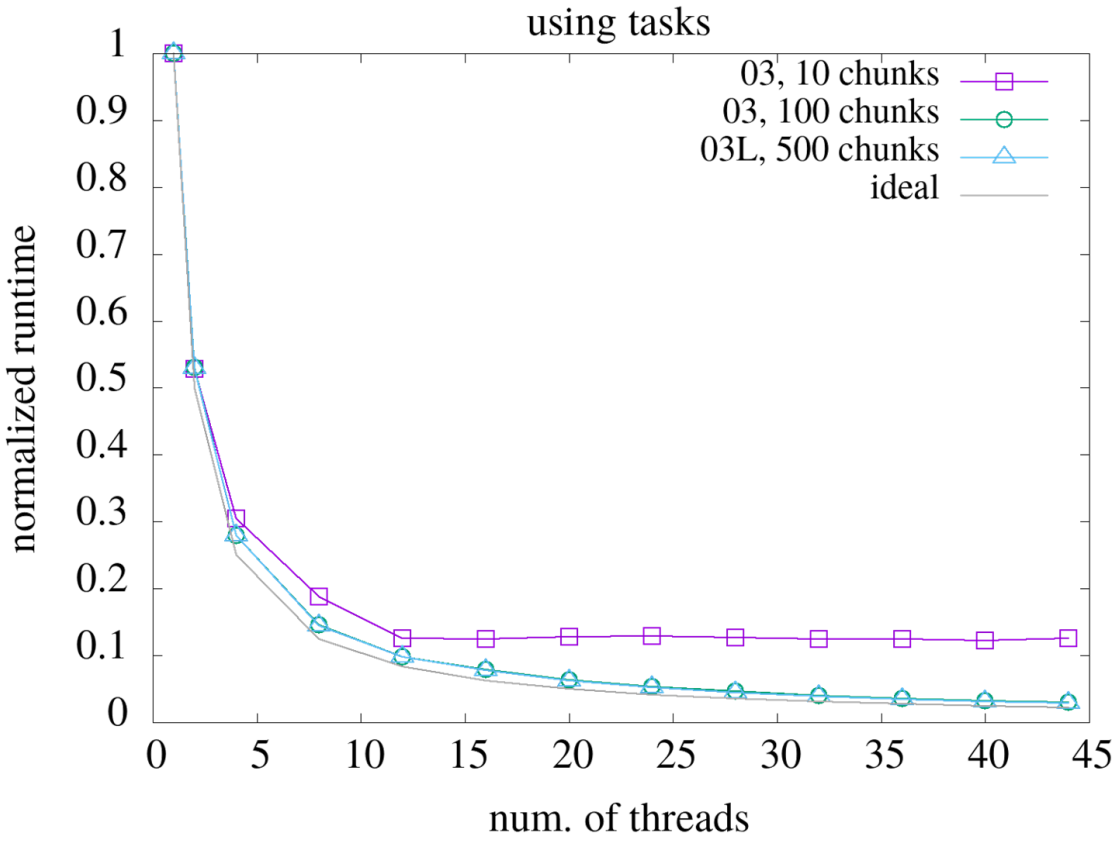
```
#pragma parallel region
{
   …;
   #pragma omp single nowait
   {
      while( !end_of_list(node) ) {
        if( node_is_to_be_processesed(node) )
           #pragma omp task
           process_node ( node );
        node = next_node( node );
   }
   …;
}
```

A task is generated for each node that must be processed

The calling thread continues traversing the linked list

Due to the `nowait` clause, all the threads skip the implied barrier at the end of the `single` region and wait here for being assigned a task

OpenMP tasks


using tasks
normalized runtime
num. of threads
03, 10 chunks
03, 100 chunks
03L, 500 chunks
ideal
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45

Hybrid
MPI + OpenMP
codes

that's all, have fun