10. Multiprocessor Scheduling (Advanced)

Operating System: Three Easy Pieces

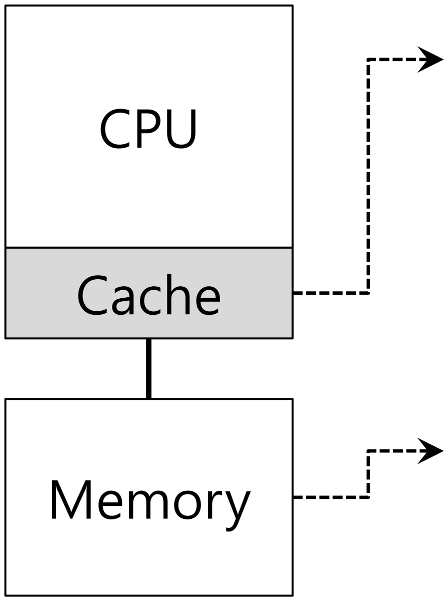
Multiprocessor Scheduling

- The rise of the multicore processor is the source of multiprocessor-scheduling proliferation.
  - **Multicore**: Multiple CPU cores are packed onto a single chip.

- Adding more CPUs does not make that single application run faster.
  → You'll have to rewrite application to run in parallel, using **threads**.

How to schedule jobs on Multiple CPUs?

Single CPU with cache



Cache

- Small, fast memories
- Hold copies of popular data that is found in the main memory.
- Utilize *temporal* and *spatial* locality

Main Memory

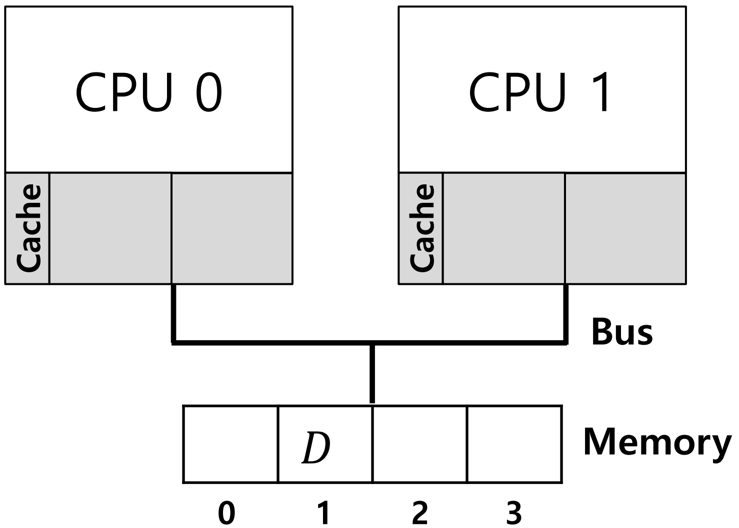
- Holds all of the data
- Access to main memory is slower than cache.

By keeping data in cache, the system can make slow memory appear to be a fast one

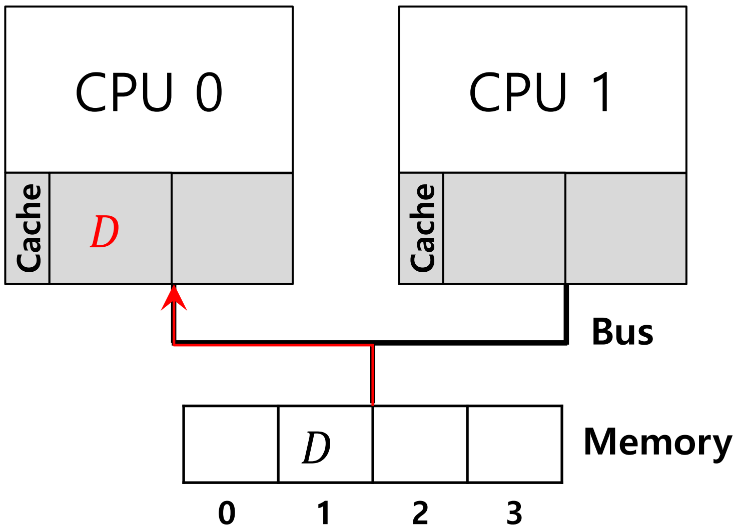
Cache coherence

- Coherence of shared resource data stored in multiple caches.

0. Two CPUs with caches sharing memory

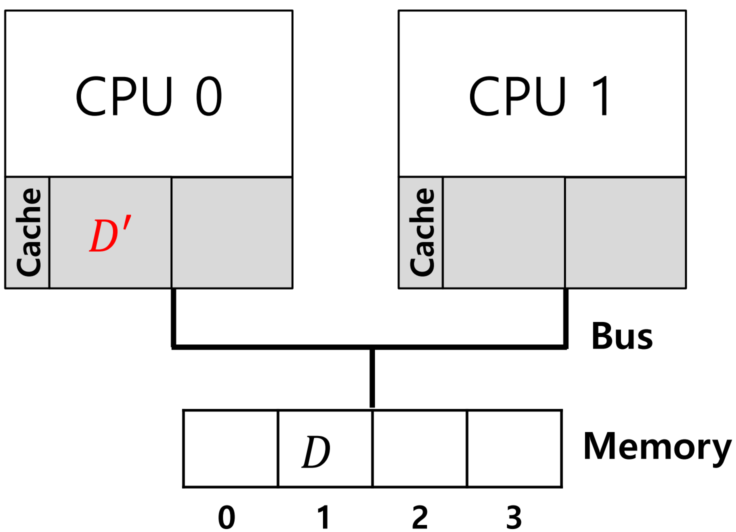


1. CPU0 reads a data at address 1.

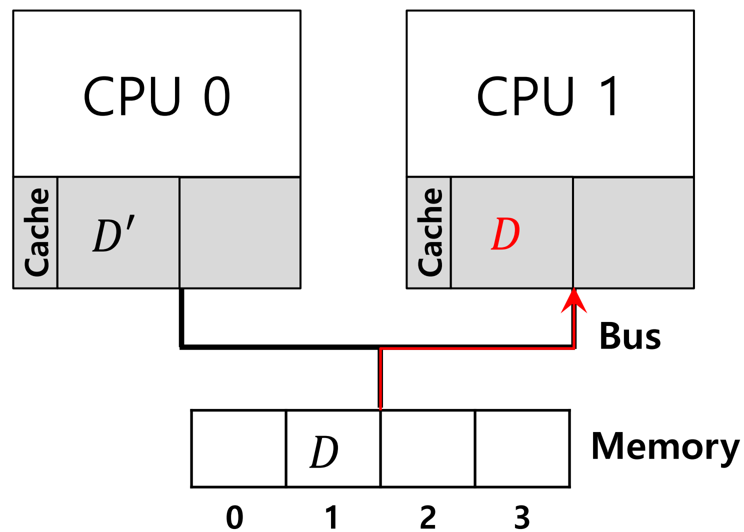


Cache coherence (Cont.)

2. D is updated and CPU1 is scheduled.



3. CPU1 re-reads the value at address A



CPU1 gets the old value D instead of the correct value D'.

Cache coherence solution

- Bus snooping
  - Each cache pays attention to memory updates by **observing the bus**.
  - When a CPU sees an update for a data item it holds in its cache, it will notice the change and either <u>invalidate</u> its copy or <u>update</u> it.

Don't forget synchronization

- When accessing shared data across CPUs, mutual exclusion primitives should likely be used to guarantee correctness.

```
typedef struct __Node_t {
        int value;
        struct __Node_t *next;
} Node_t;

int List_Pop() {
        Node_t *tmp = head;           // remember old head ...
        int value = head->value;      // ... and its value
        head = head->next;            // advance head to next pointer
        free(tmp);                    // free old head
        return value;                 // return value at head
}
```

Simple List Delete Code

Don’t forget synchronization (Cont.)

- Solution

```
pthread_mtuex_t m;
typedef struct __Node_t {
        int value;
        struct __Node_t *next;
} Node_t;

int List_Pop() {
        lock(&m)
        Node_t *tmp = head;           // remember old head ...
        int value = head->value;      // ... and its value
        head = head->next;            // advance head to next pointer
        free(tmp);                    // free old head
        unlock(&m)
        return value;                 // return value at head
}
```

Simple List Delete Code with lock

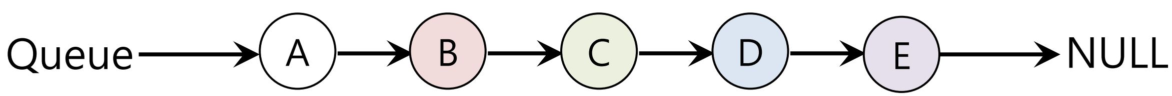
Cache Affinity

- Keep a process on the same CPU if possible
  - A process builds up a fair bit of state in the cache of a CPU.
  - The next time the process run, it will run faster if some of its state is *already present* in the cache on that CPU.

A multiprocessor scheduler should consider cache affinity when making its scheduling decision.

Single queue Multiprocessor Scheduling (SQMS)

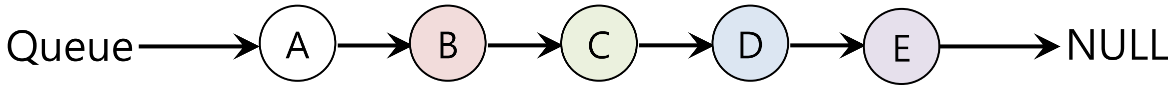
- Put all jobs that need to be scheduled into a single queue.
  - Each CPU simply picks the next job from the globally shared queue.
  - Cons:
    - Some form of **locking** must be used → Lack of scalability
    - Cache affinity
    - Example:



    - Possible job scheduler across CPUs:

CPU0	A	E	D	C	B	… (repeat) …
CPU1	B	A	E	D	C	… (repeat) …
CPU2	C	B	A	E	D	… (repeat) …
CPU3	D	C	B	A	E	… (repeat) …

Scheduling Example with Cache affinity



CPU0	A	E	A	A	A	... (repeat) ...
CPU1	B	B	E	B	B	... (repeat) ...
CPU2	C	C	C	E	C	... (repeat) ...
CPU3	D	D	D	D	E	... (repeat) ...

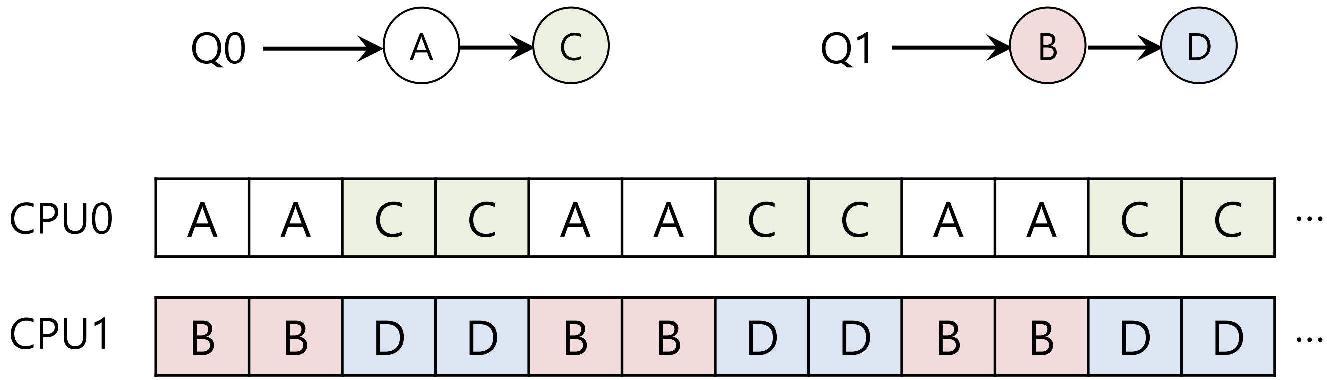
- Preserving affinity for most
  - Jobs A through D are not moved across processors.
  - Only job e Migrating from CPU to CPU.
- Implementing such a scheme can be **complex**.

Multi-queue Multiprocessor Scheduling (MQMS)

- MQMS consists of multiple scheduling queues.
  - Each queue will follow a particular scheduling discipline.
  - When a job enters the system, it is placed on **exactly one** scheduling queue.
  - Avoid the problems of information sharing and synchronization.

MQMS Example

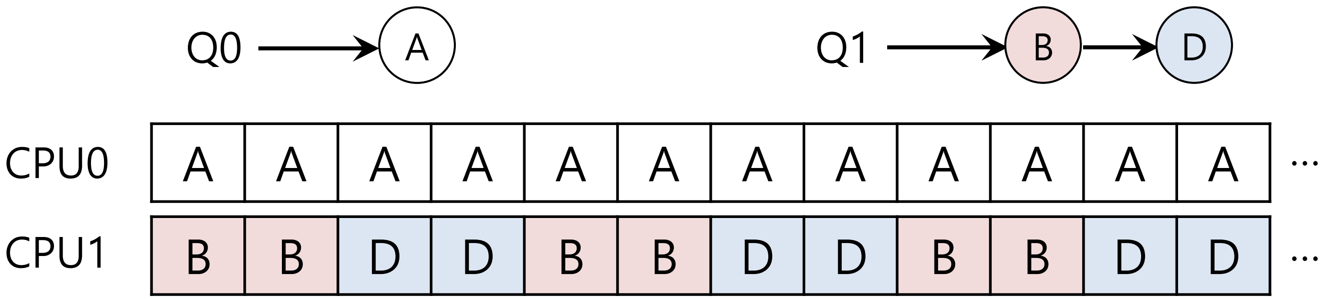
- With **round robin**, the system might produce a schedule that looks like this:



MQMS provides more scalability and cache affinity.

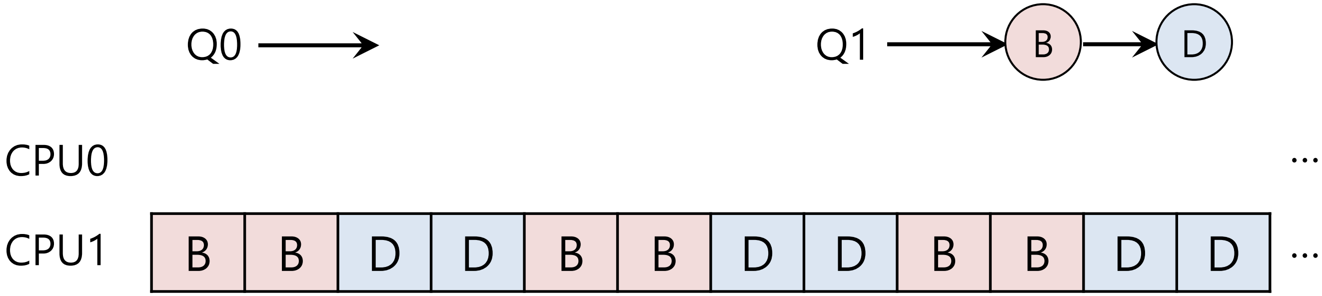
Load Imbalance issue of MQMS

- After job C in Q0 finishes:



A gets twice as much CPU as B and D.

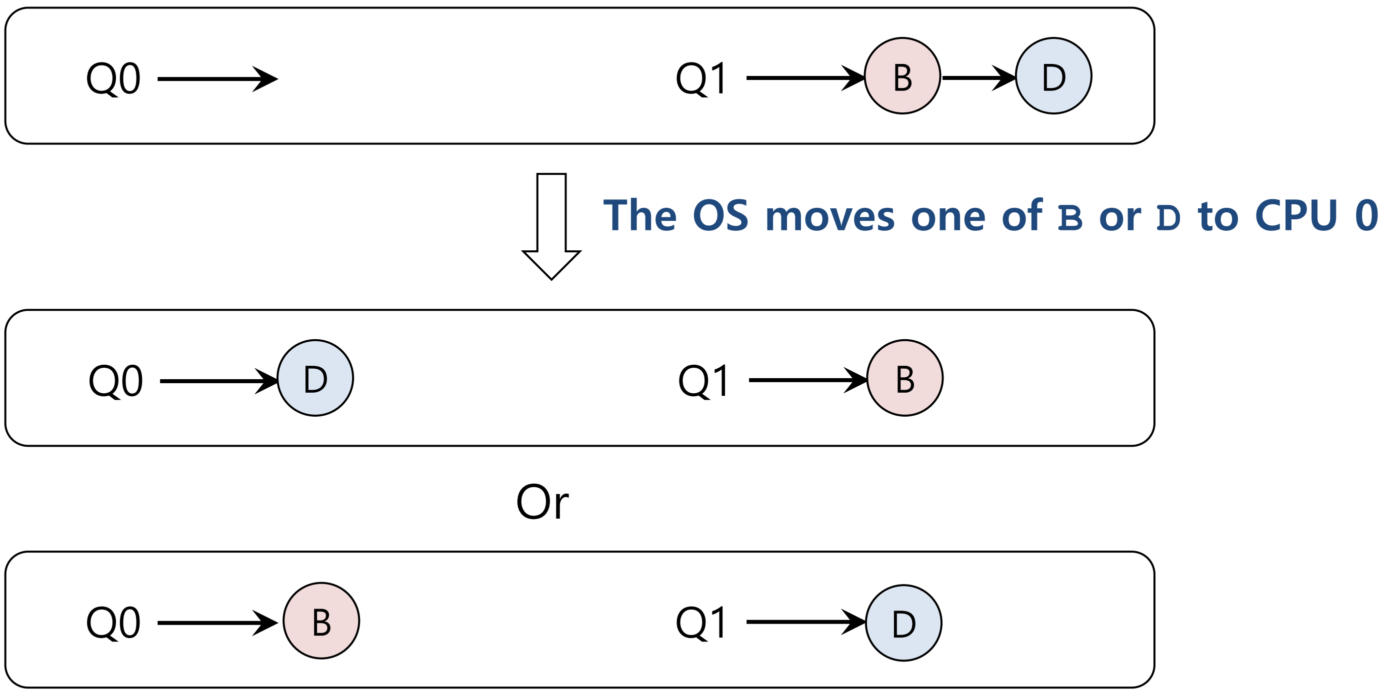
- After job A in Q0 finishes:



CPU0 will be left idle!

How to deal with load imbalance?

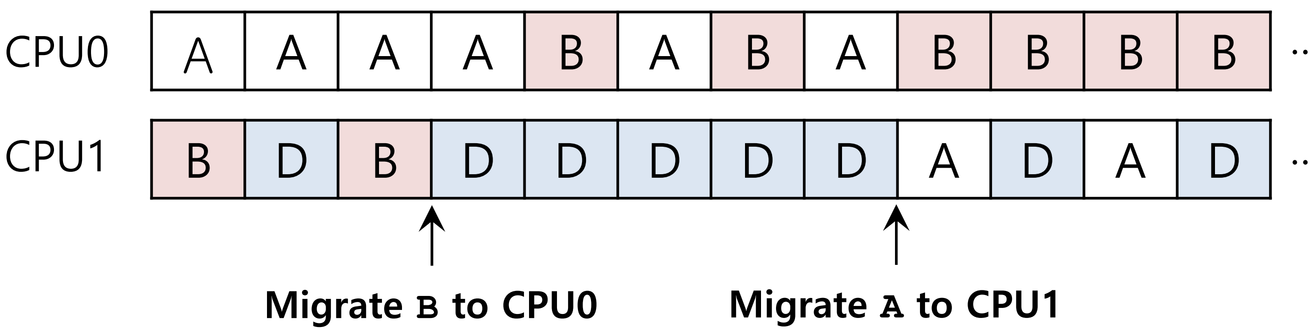
- The answer is to move jobs (**Migration**).
  - Example:



How to deal with load imbalance? (Cont.)

- A more tricky case:

- A possible migration pattern:
  - Keep switching jobs



Work Stealing

- Move jobs between queues
  - Implementation:
    - A source queue that is <u>low on jobs</u> is picked.
    - The source queue occasionally peeks at another target queue.
    - If the target queue is <u>fuller than</u> the source queue, the source will "**steal**" one or more jobs from the target queue.
  - Cons:
    - *High overhead* and trouble *scaling*

Linux Multiprocessor Schedulers

- O(1)
  - A Priority-based scheduler
  - Use Multiple queues (similar to MLFQ)
  - Change a process's priority over time
  - Schedule those with highest priority
  - Interactivity is a particular focus

- Completely Fair Scheduler (CFS) (current mainline)
  - Deterministic proportional-share approach
  - Based on Staircase Deadline (fairness is the focus)
  - Red-black tree for scalability

Linux Multiprocessor Schedulers (Cont.)

- BF Scheduler (BFS) (Not in the mainline)
  - A single queue approach
  - Proportional-share
  - Based on Earliest Eligible Virtual Deadline First (EEVDF)
  - Focus on interactive (not scale well with cores). Superseded by MuQSS to fix that

- The battle of schedulers : Kolivas (SD) vs Molnar (CFS)

"And you have to realize that there are not very many things that have a ged as well as the scheduler. Which is just another proof that scheduling is easy."

Linus Torvalds, 2001 [43]

Scheduling is not easy!, E.g:

"The Linux Scheduler: a Decade of Wasted Cores "
http://www.ece.ubc.ca/~sasha/papers/eurosys16-final29.pdf

Aside: Current hardware is hard to handle by the scheduler

- SMT, Frequency scaling,…
- Complex caches (Non-uniform cache architectures)
- Multi-socket is hard
- Multi-die is even harder
- Big-little (P/E), etc…

Aside: real systems are really nasty

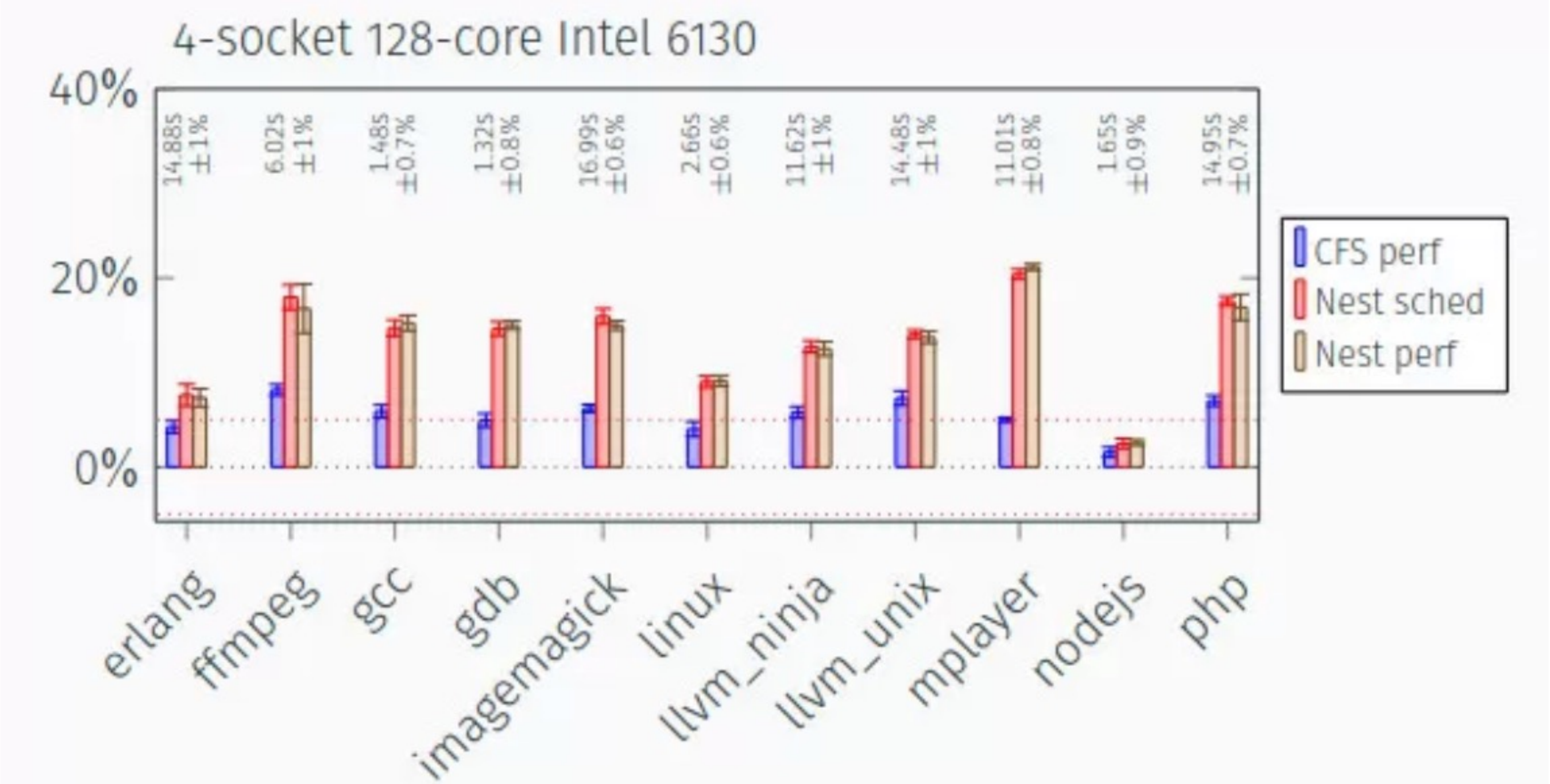
Ryzen 9 3950X Core-to-Core Latency

C -> C (ns)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	X	6.7	31.3	31.3	30.9	30.9	31.3	31.4	83.5	83.2	86.9	87.7	83.7	83.8	84.0	84.0	84.3	84.4	83.1	83.1	86.4	86.3	85.6	85.5	82.6	82.5	83.1	83.1	88.8	88.9	85.7	85.6
1	6.7	X	31.0	31.0	30.7	30.8	31.2	31.2	82.7	82.8	86.0	86.1	83.9	83.8	83.9	83.9	84.3	84.0	83.1	83.0	86.3	86.3	85.4	85.5	82.5	82.5	83.1	83.1	88.8	88.8	85.7	85.6
2	31.3	31.0	X	6.8	30.9	31.0	31.8	31.8	85.5	85.4	88.9	89.0	85.8	85.8	86.0	85.9	82.1	82.1	82.3	82.3	85.5	85.5	83.4	83.5	81.7	81.7	82.3	82.3	86.0	86.0	83.5	83.6
3	31.3	31.0	6.8	X	30.9	30.9	31.8	31.8	85.4	84.8	88.7	88.4	85.7	85.7	86.0	86.0	82.1	82.1	82.3	82.3	85.5	85.5	83.5	83.5	81.8	81.8	82.3	82.3	86.0	85.9	83.5	83.4
4	30.9	30.7	30.9	30.9	X	6.9	30.5	30.6	82.2	82.1	84.9	85.2	82.4	82.5	82.5	82.5	82.1	82.0	83.4	83.6	86.5	86.5	83.4	83.4	82.8	82.8	83.4	83.4	85.7	86.2	83.4	83.5
5	30.9	30.8	31.0	30.9	6.9	X	30.5	30.6	82.1	82.1	85.0	84.8	82.5	82.5	82.5	82.6	82.1	82.0	83.4	83.4	86.5	86.5	83.5	83.4	82.7	82.8	83.4	83.3	85.8	86.2	83.5	83.5
6	31.3	31.2	31.8	31.8	30.5	30.5	X	6.9	83.0	83.1	85.6	85.7	82.5	82.5	82.6	82.6	85.5	85.8	86.1	86.3	89.3	89.3	87.2	87.2	85.5	85.5	86.2	86.3	89.6	89.4	87.2	87.2
7	31.4	31.2	31.8	31.8	30.6	30.6	6.9	X	83.0	83.2	85.6	85.6	82.5	82.6	82.6	82.6	85.7	85.7	86.1	86.2	89.3	89.3	87.2	87.2	85.5	85.4	86.2	86.2	88.8	89.5	87.2	87.2
8	83.5	82.7	85.5	85.4	82.2	82.1	83.0	83.0	X	6.9	31.7	31.6	30.9	30.9	31.2	31.2	83.9	84.1	83.1	83.1	86.3	86.3	85.3	85.4	82.5	82.5	83.1	83.2	88.9	88.9	85.6	85.7
9	83.2	82.8	85.4	84.8	82.1	82.1	83.1	83.2	6.9	X	31.7	31.6	30.9	30.9	31.2	31.2	84.3	84.0	83.1	83.1	86.3	86.3	85.7	85.4	82.5	82.5	83.1	83.1	88.9	88.9	85.6	85.6
10	86.9	86.0	88.9	88.7	84.9	85.0	85.6	85.6	31.7	31.7	X	6.9	31.1	31.2	31.9	31.8	82.1	82.0	82.5	82.4	85.5	85.5	83.4	83.4	81.8	81.8	82.3	82.3	86.0	86.0	83.5	83.5
11	87.7	86.1	89.0	88.4	85.2	84.8	85.7	85.6	31.6	31.6	6.9	X	31.2	31.2	31.9	31.9	82.1	82.1	82.3	82.3	85.5	85.5	83.6	83.5	81.8	81.8	82.3	82.3	86.0	86.0	83.5	83.5
12	83.7	83.9	85.8	85.7	82.4	82.5	82.5	82.5	30.9	30.9	31.1	31.2	X	6.9	30.6	30.6	82.0	82.1	82.2	83.3	86.4	86.6	83.3	83.4	82.8	82.8	83.4	83.4	86.2	86.2	83.5	83.5
13	83.8	83.8	85.8	85.7	82.5	82.5	82.5	82.6	30.9	30.9	31.2	31.2	6.9	X	30.6	30.6	82.0	82.0	83.4	83.7	86.5	86.5	83.5	83.4	82.7	82.7	83.4	83.4	86.1	86.1	83.5	83.5
14	84.0	83.9	86.0	86.0	82.5	82.5	82.6	82.6	31.2	31.2	31.9	31.9	30.6	30.6	X	6.9	85.8	85.7	86.0	86.2	89.3	89.3	87.2	87.2	85.5	85.5	86.2	86.2	88.7	89.5	87.2	87.2
15	84.0	83.9	85.9	86.0	82.5	82.6	82.6	82.6	31.2	31.2	31.8	31.9	30.6	30.6	6.9	X	85.7	85.7	86.1	86.2	89.3	89.3	87.2	87.2	85.5	85.5	86.2	86.2	88.5	89.5	87.2	87.2
16	84.3	84.3	82.1	82.1	82.1	82.1	85.5	85.7	83.9	84.3	82.1	82.1	82.0	82.0	85.8	85.7	X	7.2	33.1	33.1	32.1	32.1	32.6	32.5	84.9	84.9	85.1	85.1	88.6	88.6	86.1	86.2
17	84.4	84.0	82.1	82.1	82.0	82.0	85.8	85.7	84.1	84.0	82.0	82.1	82.1	82.0	85.7	85.7	7.2	X	33.1	33.1	32.1	32.1	32.5	32.5	84.9	84.8	85.1	85.1	88.6	88.6	86.1	86.2
18	83.1	83.1	82.3	82.3	83.4	83.4	86.1	86.1	83.1	83.1	82.5	82.3	82.2	83.4	86.0	86.1	33.1	33.1	X	7.1	32.3	32.4	33.0	33.0	85.0	85.0	85.4	85.4	88.8	89.0	86.5	86.6
19	83.1	83.0	82.3	82.3	83.6	83.4	86.3	86.2	83.1	83.1	82.4	82.3	83.3	83.7	86.2	86.2	33.1	33.1	7.1	X	32.4	32.4	33.0	33.1	85.0	85.1	85.4	85.4	89.0	88.9	86.5	86.5
20	86.4	86.3	85.5	85.5	86.5	86.5	89.3	89.3	86.3	86.3	85.5	85.5	86.4	86.5	89.3	89.3	32.1	32.1	32.3	32.4	X	7.1	31.9	31.9	88.3	88.3	88.6	88.6	92.2	92.3	90.0	90.1
21	86.3	86.3	85.5	85.5	86.5	86.5	89.3	89.3	86.3	86.3	85.5	85.5	86.6	86.5	89.3	89.3	32.1	32.1	32.4	32.4	7.1	X	31.9	32.0	88.3	88.3	88.6	88.6	92.5	92.4	90.0	90.1
22	85.6	85.4	83.4	83.5	83.4	83.5	87.2	87.2	85.3	85.7	83.4	83.6	83.3	83.5	87.2	87.2	32.6	32.5	33.0	33.0	31.9	31.9	X	7.2	86.2	86.1	86.4	86.4	90.0	90.1	87.5	87.5
23	85.5	85.5	83.5	83.5	83.4	83.4	87.2	87.2	85.4	85.4	83.4	83.5	83.4	83.4	87.2	87.2	32.5	32.5	33.0	33.1	31.9	32.0	7.2	X	86.2	86.2	86.5	86.4	90.0	90.1	87.6	87.6
24	82.6	82.5	81.7	81.8	82.8	82.7	85.5	85.5	82.5	82.5	81.8	81.8	82.8	82.7	85.5	85.5	84.9	84.9	85.0	85.0	88.3	88.3	86.2	86.2	X	7.1	32.9	32.9	32.1	32.1	32.6	32.6
25	82.5	82.5	81.7	81.8	82.8	82.8	85.5	85.4	82.5	82.5	81.8	81.8	82.8	82.7	85.5	85.5	84.9	84.8	85.0	85.1	88.3	88.3	86.1	86.2	7.1	X	32.9	32.9	32.1	32.1	32.6	32.6
26	83.1	83.1	82.3	82.3	83.4	83.4	86.2	86.2	83.1	83.1	82.3	82.3	83.4	83.4	86.2	86.2	85.1	85.1	85.4	85.4	88.6	88.6	86.4	86.5	32.9	32.9	X	7.1	32.5	32.5	33.1	33.1
27	83.1	83.1	82.3	82.3	83.4	83.3	86.3	86.2	83.2	83.1	82.3	82.3	83.4	83.4	86.2	86.2	85.1	85.1	85.4	85.4	88.6	88.6	86.4	86.4	32.9	32.9	7.1	X	32.6	32.5	33.1	33.1
28	88.8	88.8	86.0	86.0	85.7	85.8	89.6	88.8	88.9	88.9	86.0	86.0	86.2	86.1	88.7	88.5	88.6	88.6	88.8	89.0	92.2	92.5	90.0	90.0	32.1	32.1	32.5	32.6	X	7.2	31.9	32.0
29	88.9	88.8	86.0	85.9	86.2	86.2	89.4	89.5	88.9	88.9	86.0	86.0	86.2	86.1	89.5	89.5	88.6	88.6	89.0	88.9	92.3	92.4	90.1	90.1	32.1	32.1	32.5	32.5	7.2	X	31.9	32.0
30	85.7	85.7	83.5	83.5	83.4	83.5	87.2	87.2	85.6	85.6	83.5	83.5	83.5	83.5	87.2	87.2	86.1	86.1	86.5	86.5	90.0	90.0	87.5	87.6	32.6	32.6	33.1	33.1	31.9	31.9	X	7.2
31	85.6	85.6	83.6	83.4	83.5	83.5	87.2	87.2	85.7	85.6	83.5	83.5	83.5	83.5	87.2	87.2	86.2	86.2	86.6	86.5	90.1	90.1	87.5	87.6	32.6	32.6	33.1	33.1	32.0	32.0	7.2	X

```
vpuente@compute-gpu-0:~$ sudo numactl --hardware
available: 8 nodes (0-7)
node 0 cpus: 0 1 2 3 32 33 34 35
node 0 size: 32097 MB
node 0 free: 30877 MB
node 1 cpus: 4 5 6 7 36 37 38 39
node 1 size: 16125 MB
node 1 free: 15219 MB
node 2 cpus: 8 9 10 11 40 41 42 43
node 2 size: 16125 MB
node 2 free: 14937 MB
node 3 cpus: 12 13 14 15 44 45 46 47
node 3 size: 8061 MB
node 3 free: 6498 MB
node 4 cpus: 16 17 18 19 48 49 50 51
node 4 size: 32253 MB
node 4 free: 31868 MB
node 5 cpus: 20 21 22 23 52 53 54 55
node 5 size: 16125 MB
node 5 free: 10104 MB
node 6 cpus: 24 25 26 27 56 57 58 59
node 6 size: 16104 MB
node 6 free: 8043 MB
node 7 cpus: 28 29 30 31 60 61 62 63
node 7 size: 8060 MB
node 7 free: 7271 MB
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  16  16  16  32  32  32  32
  1:  16  10  16  16  32  32  32  32
  2:  16  16  10  16  32  32  32  32
  3:  16  16  16  10  32  32  32  32
  4:  32  32  32  32  10  16  16  16
  5:  32  32  32  32  16  10  16  16
  6:  32  32  32  32  16  16  10  16
  7:  32  32  32  32  16  16  16  10
```

Aaside: Nest Scheduler

- https://www.phoronix.com/news/Nest-Linux-Scheduling-Warm-Core
- Idea: Form a primary and secondary nest to choose the core to run
  - Locality, prefer warm-cores (higher frequency)



Aside: Other Solutions

- Strawman solution:
  - `taskset –p <range-processors> your_task`

- Hardware assistance?
  - V.gr., Intel Thread Director (Alder Lake + windows 11, Linux 5.18)
  - https://www.anandtech.com/show/16881/a-deep-dive-into-intels-alder-lake-microarchitectures/2

- Disclaimer: Disclaimer: This lecture slide set is used in AOS course at University of Cantabria. Was initially developed for Operating System course in Computer Science Dept. at Hanyang University. This lecture slide set is for OSTEP book written by Remzi and Andrea Arpaci-Dusseau (at University of Wisconsin)