CE4302 – Arquitectura de Computadores II

Multiprocesamiento en chip (CMP)

PROFESOR: ING. LUIS BARBOZA ARTAVIA

Agenda

- Introducción
- CMP
- Retos en el paralelismo

Investigar

1- Qué es *thread level speculation* (***TLS***)?

2- Que son algoritmos irregulares en el contexto de ***parallel computing***.

*3- Qué herramientas existen en la industria para analizar el **TLP**.*

*4- Investigue sobre los siguientes términos**: logical core, physical core, risc-v hart***

5- Sección 5.1 John L Hennessy y David A Patterson. Computer architecture: a quantitative approach. Elsevier, 2017.

6- Investigue que es el thread director en las nuevas architecturas de Intel y como se relaciona con el OS

Converting Thread-Level Parallelism to Instruction-Level Parallelism via Simultaneous Multithreading

JACK L. LO and SUSAN J. EGGERS
University of Washington
JOEL S. EMER
Digital Equipment Corporation
HENRY M. LEVY
University of Washington
REBECCA L. STAMM
Digital Equipment Corporation
and
DEAN M. TULLSEN
University of California, San Diego

Investigar

1- Qué es *thread level speculation* (***TLS***)?

A Survey on Thread-Level Speculation Techniques

ALVARO ESTEBANEZ, DIEGO R. LLANOS, and ARTURO GONZALEZ-ESCRIBANO,
Universidad de Valladolid

*"…Thread-Level Speculation (TLS), also called Speculative Parallelization (SP), or even Optimistic Parallelization**, is a runtime technique that executes in parallel fragments of code that were originally intended to run sequentially**…"*

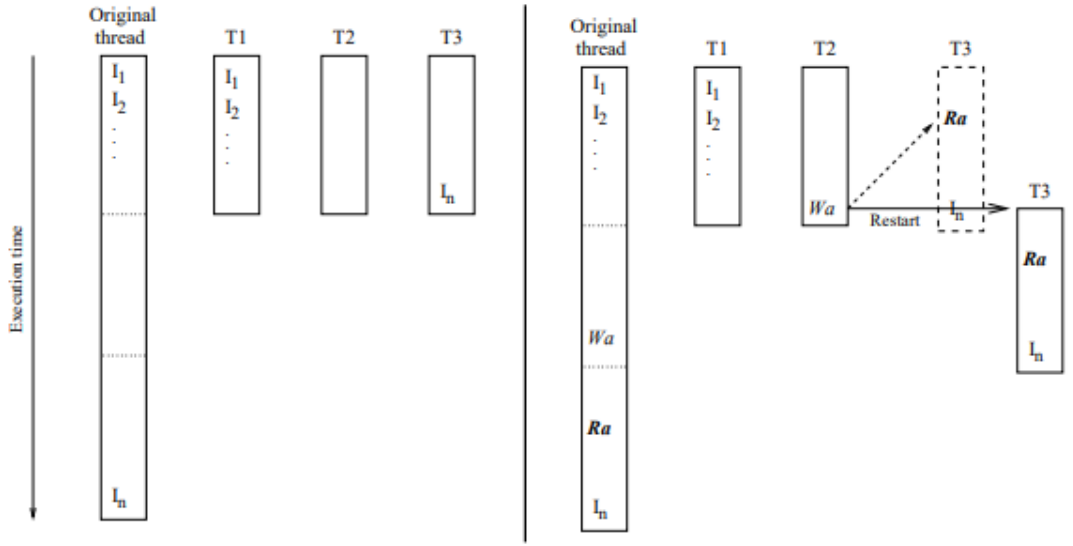


Figure 1.3: *Example of thread-level speculation (left) and a dependence violation (right).*

Techniques to Reduce Thread-Level Speculation Overhead

June 2006

Thesis for: Doctor of Philosophy

Authors:

Fredrik Warg
RISE Research Institutes of Sweden

Investigar

2- Que son algoritmos irregulares en el contexto de ***parallel computing***.

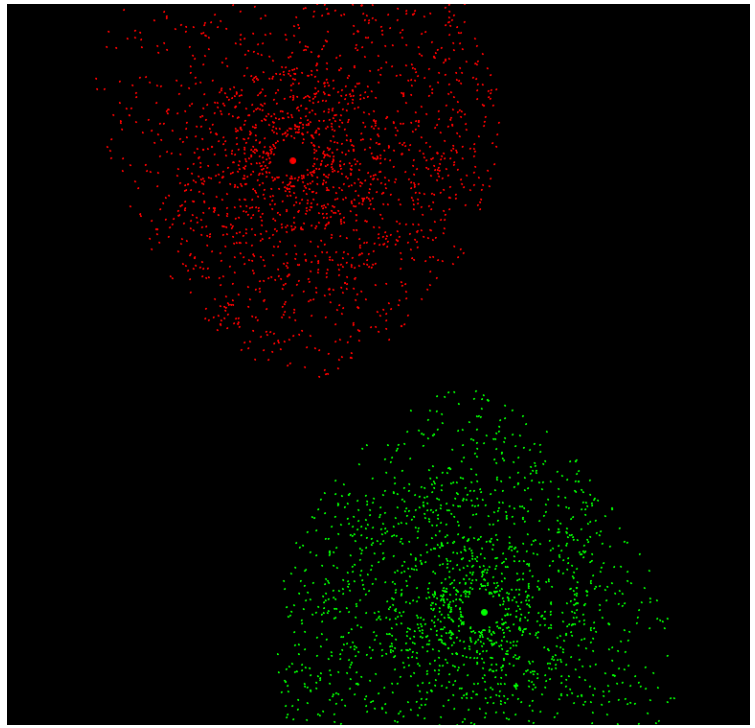
"…Irregular algorithms are a class of algorithms ***that exhibit non-uniform data access patterns and dependencies****…"*

Parallel Computing Strategies
for Irregular Algorithms

RUPAK BISWAS
NASA Ames Research Center

LEONID OLIKER and HONGZHANG SHAN
Lawrence Berkeley National Laboratory

https://ntrs.nasa.gov/api/citations/20020090950/downloads/20020090950.pdf



https://github.com/lkell/BarnesHut

Investigar

*3- Qué herramientas existen en la industria para analizar el **TLP**.*

https://www.youtube.com/watch?v=t194aY6rj1U

Threading

Analysis Configuration | Collection Log | Summary | Bottom-up | Caller/Callee | Top-down Tree | Platform

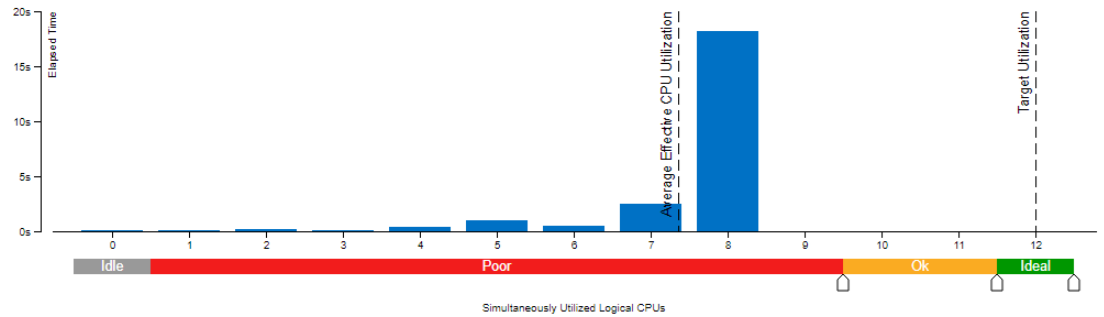
Elapsed Time: 23.034s

Paused Time: 0s

Effective CPU Utilization: 61.4% (7.364 out of 12 logical CPUs)

Effective CPU Utilization Histogram

This histogram displays a percentage of the wall time the specific number of CPUs were running simultaneously. Spin and Overhead time adds to the Idle CPU utilization value.



Total Thread Count: 9

Thread Oversubscription: 0s (0.0% of CPU Time)

Wait Time with poor CPU Utilization: 22.915s (100.0% of Wait Time)

Top Waiting Objects

This section lists the objects that spent the most time waiting in your application. Objects can wait on specific calls, such as sleep() or I/O, or on contended synchronizations. A significant amount of Wait time associated with a synchronization object reflects high contention for that object and, thus, reduced parallelism.

Sync Object	Wait Time with poor CPU Utilization	(% from Object Wait Time)	Wait Count
Thread 0xc94cf012	22.914s	100.0%	1
Stream 0xeecdd4ee	0.001s	100.0%	12

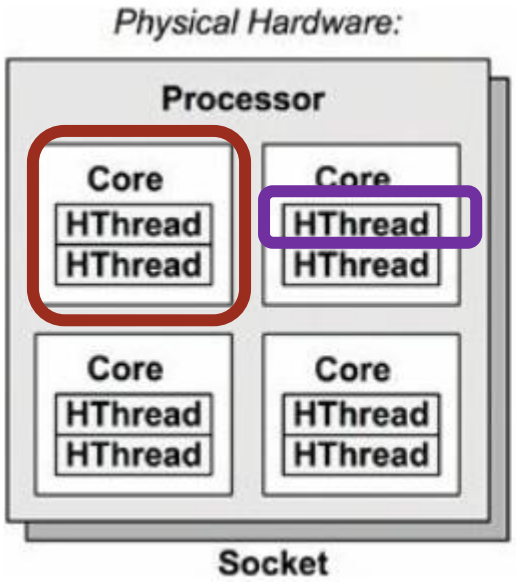
**N/A is applied to non-summable metrics.*

Spin and Overhead Time: 0.034s (0.0% of CPU Time)

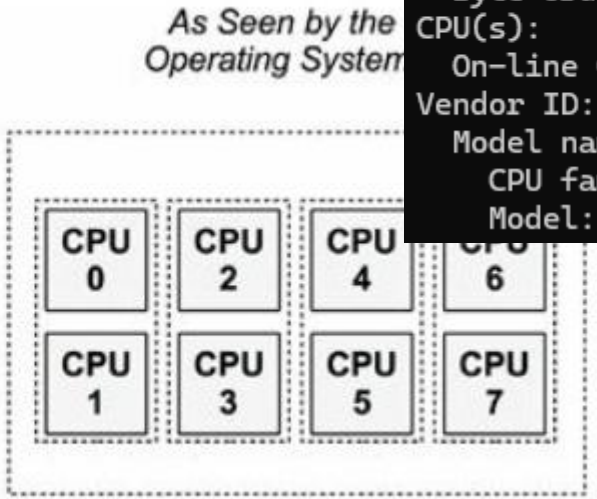
https://www.intel.com/content/www/us/en/docs/vtune-profiler/user-guide/2023-0/threading-efficiency-view.html

Investigar

4- Investigue sobre los siguientes términos: ***logical core, physical core,*** ***risc-v hart***



https://www.linkedin.com/pulse/understanding-physical-logical-cpus-akshay-deshpande/




```
(base) luis@DESKTOP-5S6FSC7:~$ lscpu
Architecture:                    x86_64
  CPU op-mode(s):                32-bit, 64-bit
  Address sizes:                 39 bits physical, 48 bits virtual
  Byte Order:                    Little Endian
CPU(s):                          16
  On-line CPU(s) list:           0-15
Vendor ID:                       GenuineIntel
  Model name:                    11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz
    CPU family:                  6
    Model:                       141
```

Base speed:	2.50 GHz
Sockets:	1
Cores:	8
Logical processors:	16
Virtualization:	Enabled

```
1   hart: A hardware execution context, which contains all the state mandated by
2   the RISC-V ISA: a PC and some registers.  This terminology is designed to
3   disambiguate software's view of execution contexts from any particular
4   microarchitectural implementation strategy.
```

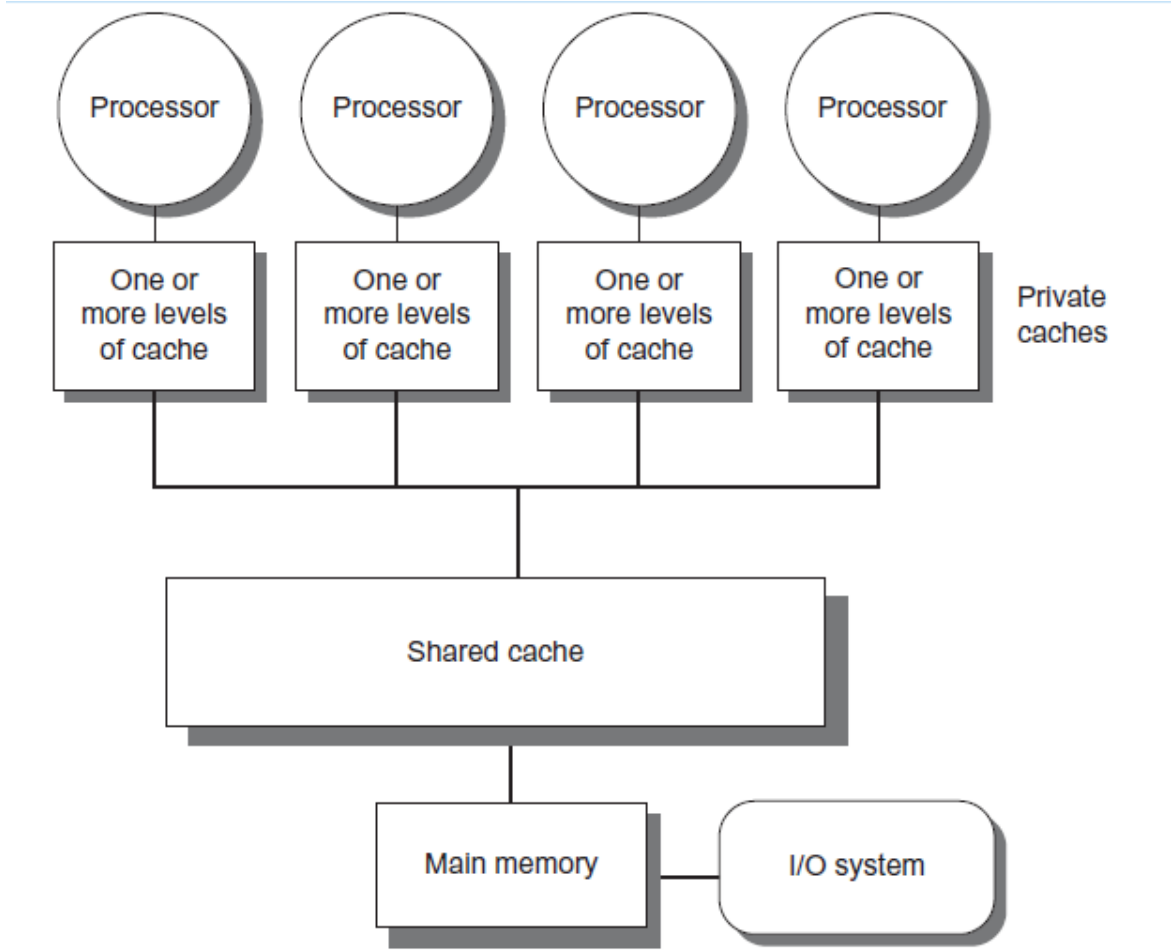
Multihilo

Sucede cuando el flujo de instrucciones se divide en varios flujos más pequeños (hilos) de manera que los hilos se puedan ejecutar de manera paralela.

Multihilo

- Multihilo intercalado (fine-grained multithreading)
- Multihilo bloqueado (coarse-grained multithreading)
- Multihilo simultáneo (SMT)
- Multiprocesamiento en chip

Multiprocesamiento en chip



Multiprocesamiento en chip

- Se implementan múltiples núcleos en un único chip y cada núcleo maneja diferentes hilos.

Características:

- Arquitecturas MIMD.
- Aplicación más común de TLP: Sistemas embebidos, propósito general, servidores, aplicaciones científicas, etc.

Multiprocesamiento en chip

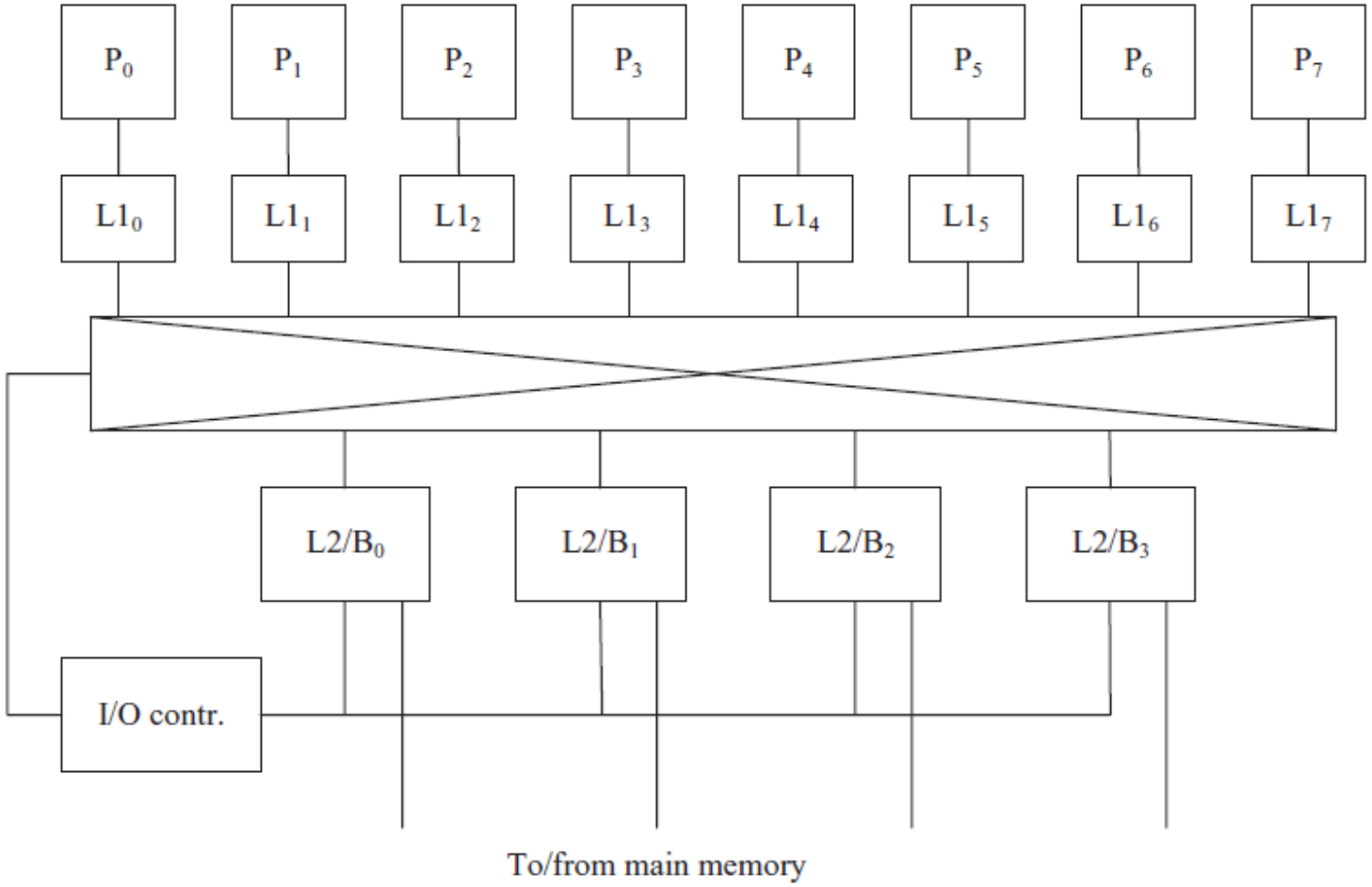
Características:

- Cantidad de multihilo.
- Puede ser en orden o fuera de orden.
- La homogeneidad del procesador.
- Jerarquía de memoria caché.
- Interconexión entre el procesador y caché.

The Sun Niagara Multiprocessor

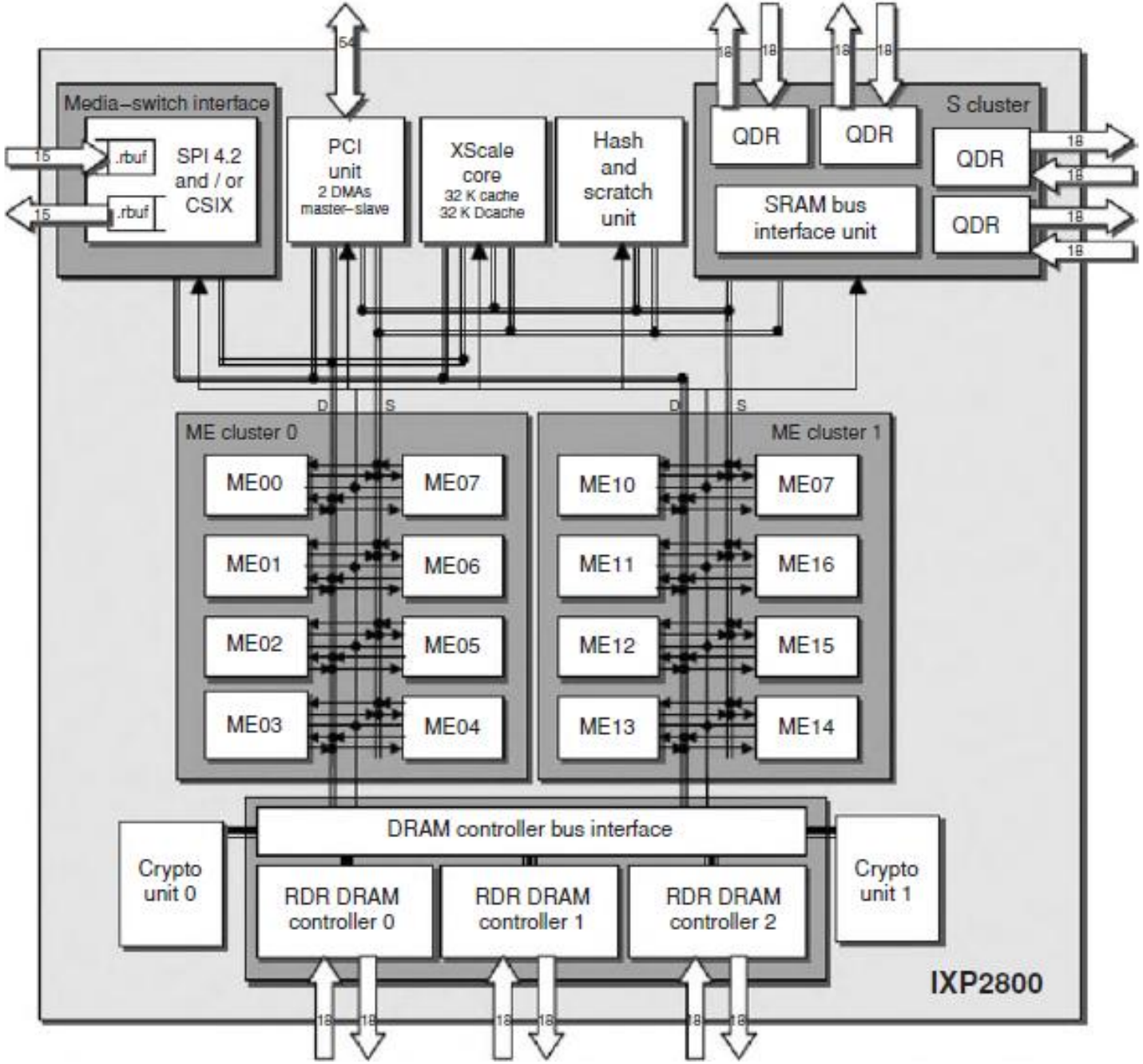
- Sun Microsystems.
- 32 hilos pueden estar activos.
- No requiere alta tasa de computación.
- Servidores web y aplicaciones con base de datos.
- Pipeline con 6 etapas.
  - Entre IF e ID.

The Sun Niagara Multiprocessor



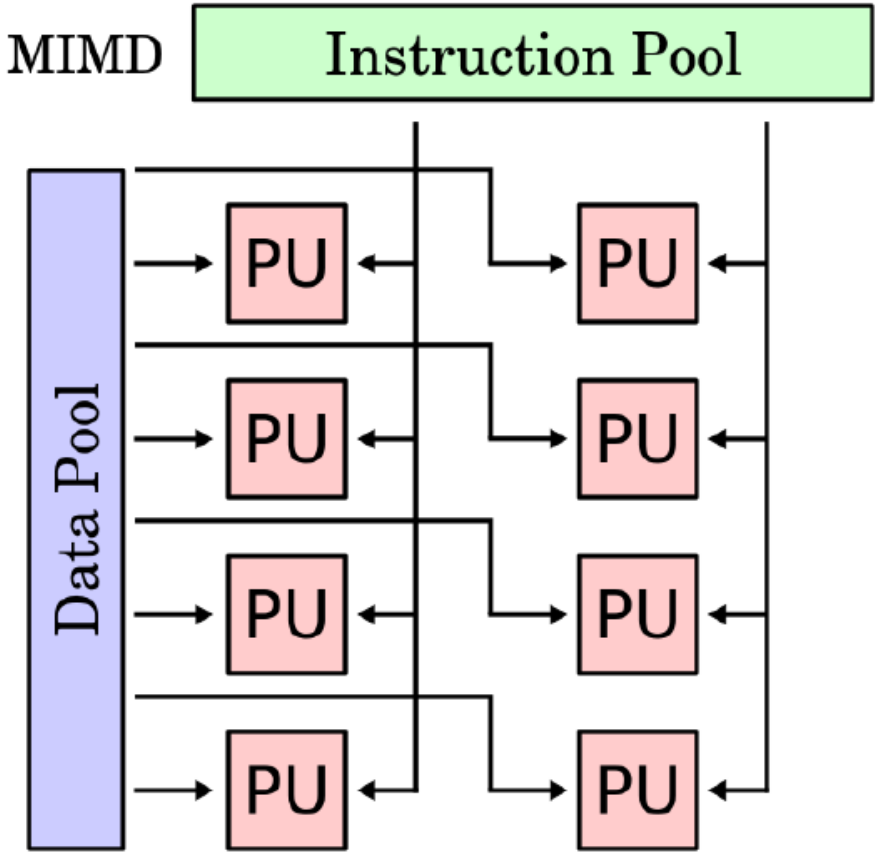
Intel IXP 2800

- Procesador de redes.
- La red ha crecido exponencialmente.
- Necesidad de aplicaciones flexibles.
- Se utiliza bloqueado iniciado desde el usuario para accesos a la memoria externa.

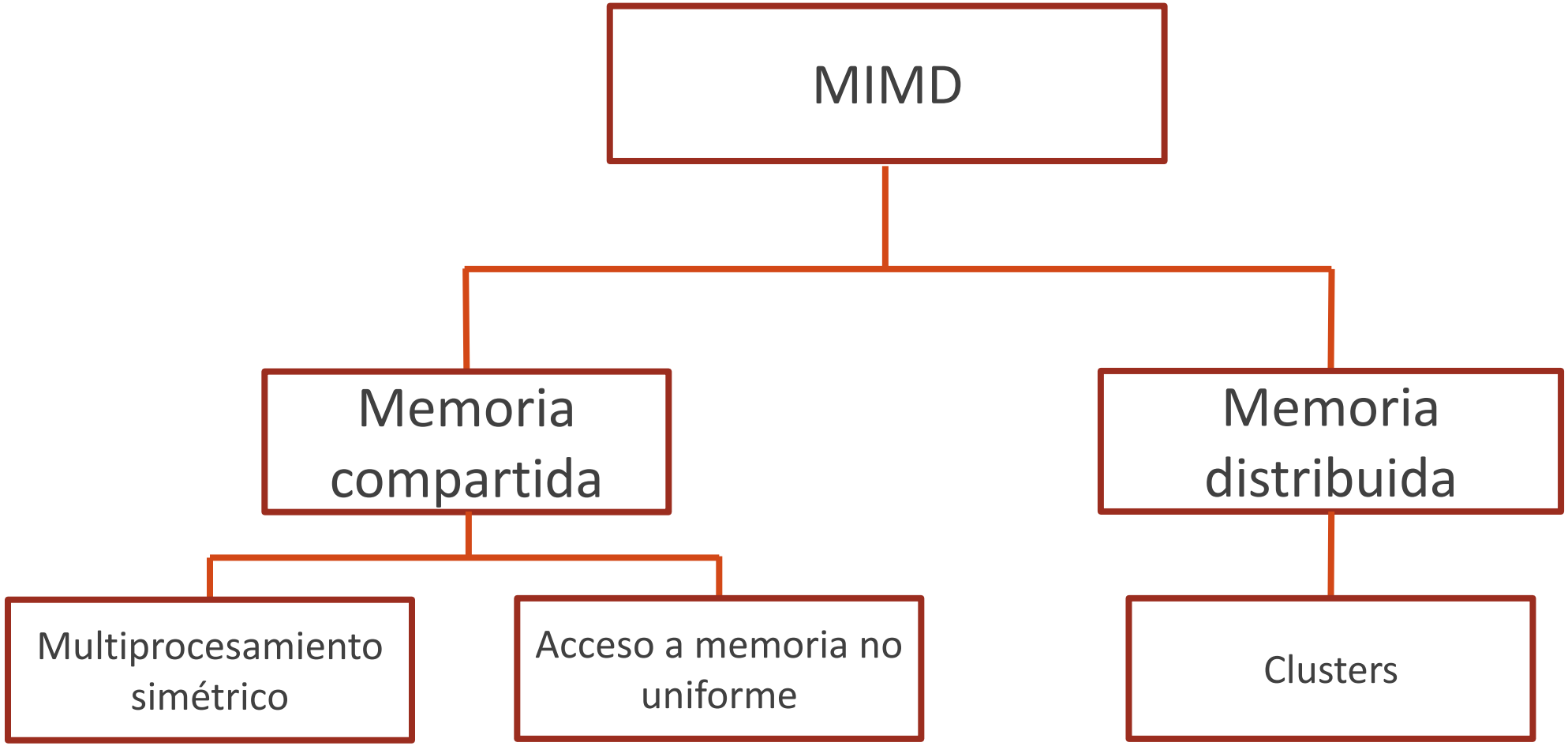


- Microengines son tareas cortas, estáticos y dinámicos.
- Paralelismo de paquetes.
- Accesos rápidos de memoria.

Multiprocesamiento en chip



Taxonomía de procesadores



Multiprocesadores simétricos (SMP)

- SO, de forma dinámica, determina los roles de cada core.
- Cada core en el cluster tiene la misma vista de memoria y HW compartido. UMA.
- Cualquier aplicación, proceso o tarea puede correr en cualquier core.
- Aplicación multihilo puede correr en varios cores a la vez.
- El SO puede esconder mucha de la complejidad de las aplicaciones.

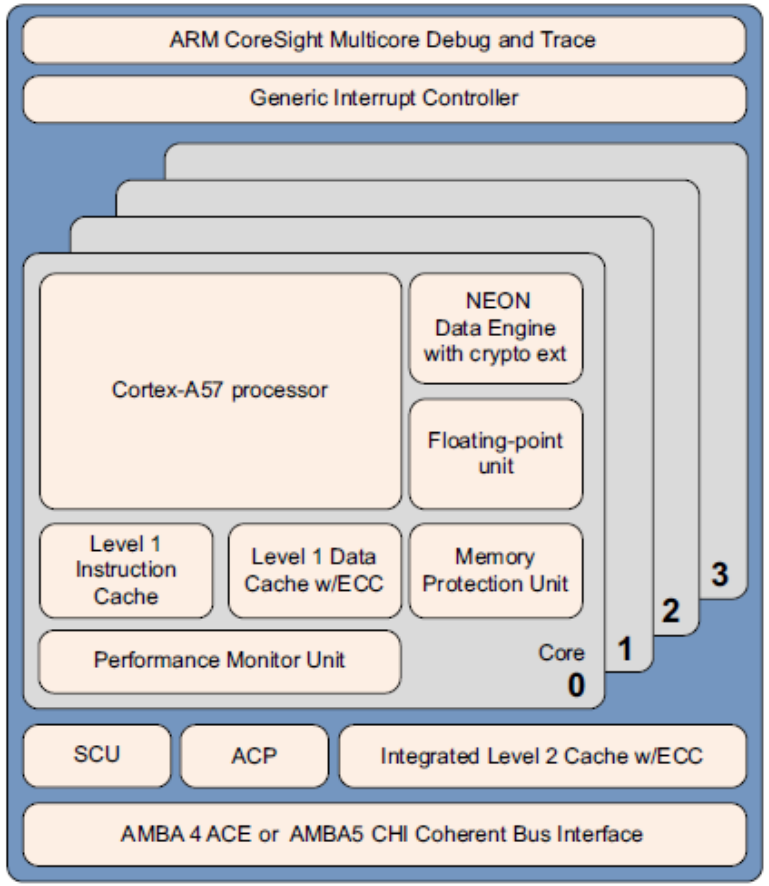


Figure 2-3 Cortex-A57 processor core

Multiprocesadores asimétricos (AMP)

- De forma estática se asignan roles individuales a los cores.
- Cada core corre un SO.
- Cada tarea puede tener una visión diferente de memoria.
- No hay requerimiento de coherencia de cache.
- Uso típico: seguridad.

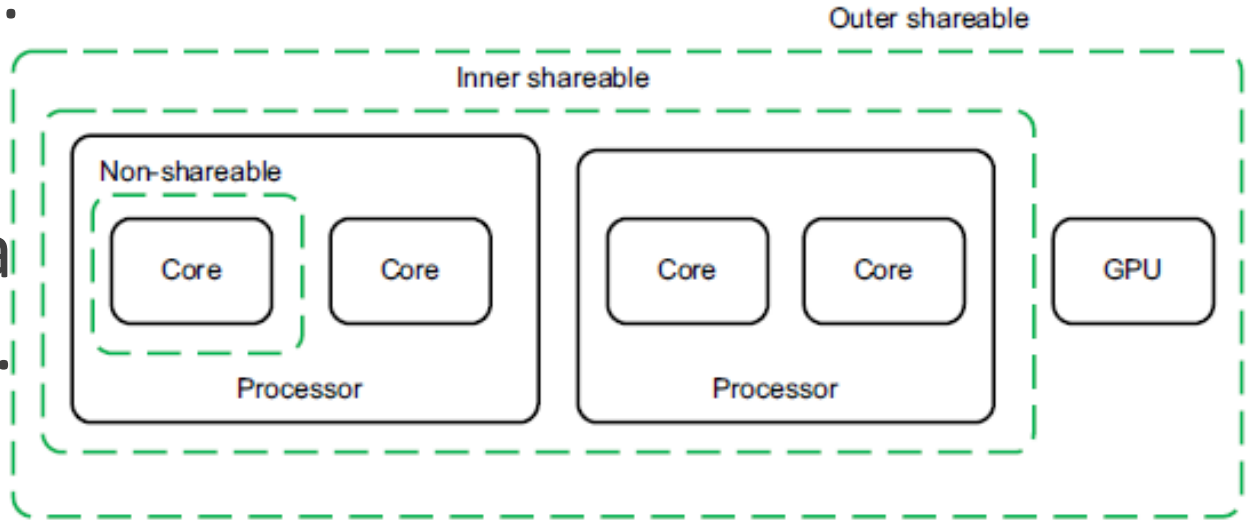
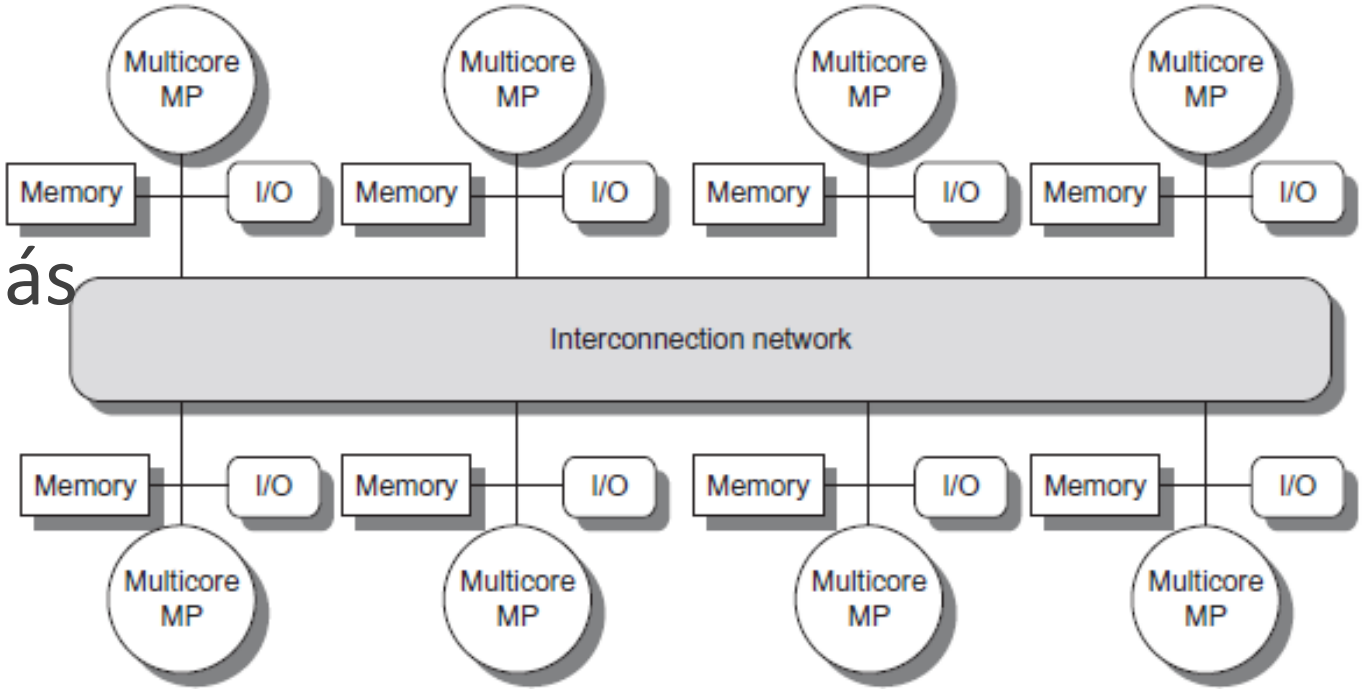


Figure 13-4 Inner and outer shareable domains

Memoria compartida distribuida (DSM)

- Memoria física distribuida.
- Tiene acceso a memoria no uniforme (NUMA).
- Comunicación entre procesadores se vuelve más compleja.



Retos en procesamiento paralelo

- **Limitación** de paralelismo en programa.
- **Comunicación:** los hilos se pueden requerir comunicarse entre sí.

Referencias

- Stallings, W. (2003). Computer organization and architecture: designing for performance. Pearson Education India.
- Hennessy, J. L., & Patterson, D. A. (2011). Computer architecture: a quantitative approach. Elsevier.

CE4302 – Arquitectura de Computadores II

Multiprocesamiento en chip (CMP)

PROFESOR: ING. LUIS BARBOZA ARTAVIA