La Arquitectura de los Chips de Alto Rendimiento

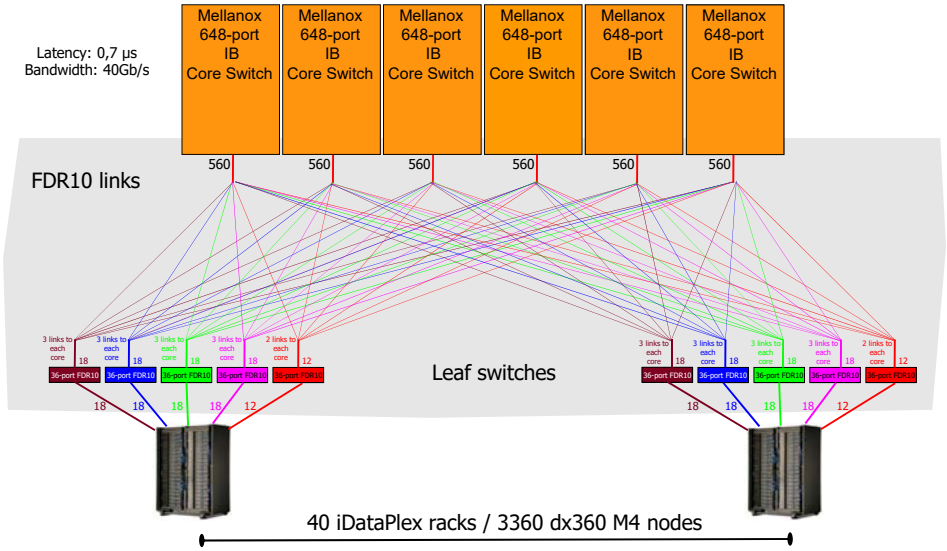
Diciembre, 2017
Máster Universitario en Data Science

Ramón Beivide
Universidad de Cantabria

Outline

- 1. Programs, Performance & Power
- 2. ISA Execution Models
- 3. Memory Basis
- 4. On-chip networks (NoCs)

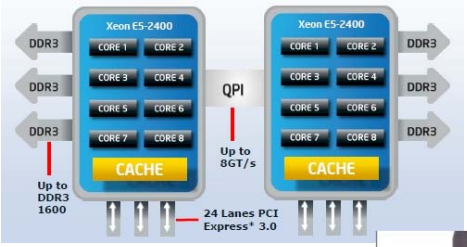
MareNostrum BSC as an example (up to 40 racks)

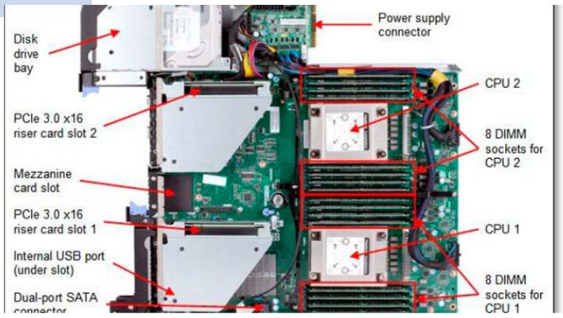


CANTABRIA CAMPUS INTERNACIONAL

3

Inside the Servers

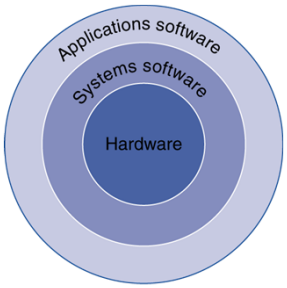




CANTABRIA CAMPUS INTERNACIONAL

4

Below Your Program



Application software

- Written in high-level language

System software

- Compiler: translates HLL code to machine code
- Operating System: service code
  - Handling input/output
  - Managing memory and storage
  - Scheduling tasks & sharing resources

Hardware

- Processor, memory, I/O controllers

Levels of Program Code

High-level language

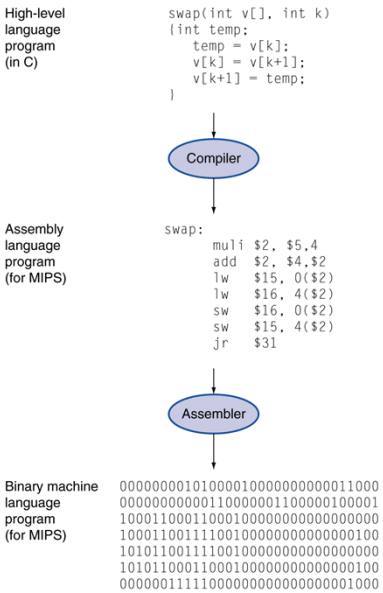
- Level of abstraction closer to problem domain
- Provides for productivity and portability

Assembly language

- Textual representation of instructions

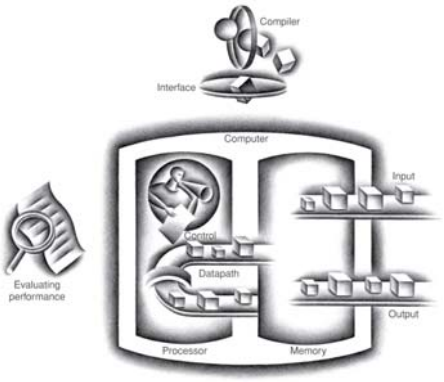
Hardware representation

- Binary digits (bits)
- Encoded instructions and data



Components of a Computer

The BIG Picture



Same components for all kinds of computer

- Desktop, server, embedded

Input/output includes

- User-interface devices
  - Display, keyboard, mouse
- Storage devices
  - Hard disk, CD/DVD, flash
- Network adapters
  - For communicating with other computers

Response Time and Throughput

Response time or latency

- How long it takes to do a task

Throughput

- Total work done per unit time
  - e.g., tasks/transactions/… per hour

How are response time and throughput affected by

- Replacing the processor with a faster version?
- Adding more processors?

Measuring Execution Time

Elapsed time

- Total response time, including all aspects
  - Processing, I/O, OS overhead, idle time
- Determines system performance

CPU time

- Time spent processing a given job
  - Discounts I/O time, other jobs' shares
- Comprises user CPU time and system CPU time
- Different programs are affected differently by CPU and system performance

Benchmarks & Workloads

Relative Performance

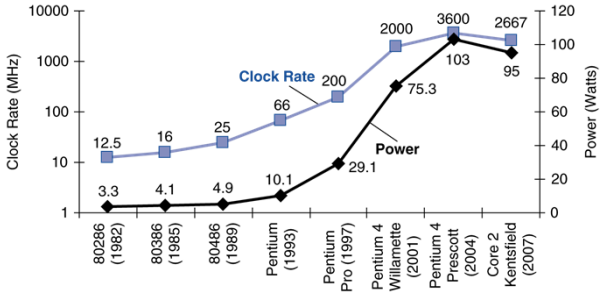
Performance = 1/Execution Time

"X is n time faster than Y"

$$\text{Performance}_X / \text{Performance}_Y = \text{Execution time}_Y / \text{Execution time}_X = n$$

- Example: time taken to run a program
  - 10s on A, 15s on B
  - Execution Time_B / Execution Time_A = 15s / 10s = 1.5
  - So A is 1.5 times faster than B

Power Trends



In CMOS IC technology

$$\text{Execution Time} = \text{Number of instructions} \times \text{CPI} \times 1/\text{Frequency}$$

$$\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}$$

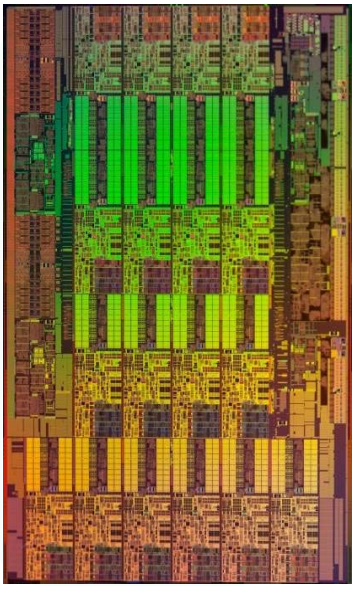
×30 | 5V → 1V | ×1000

Intel Haswell-EP Xeon E5-26xx v3 (1st Q2014)

The flagship
Xeon E5-2699 v3
which boasts
18 Cores,
36 threads,
45 MB of L3 cache and a
high price of $4115 US
(Feb. 2015)

Intel Xeon E5-2600 V3 "Haswell-EP" Workstation and Server Processors Unleashed For High-Performance Computing, Hassan Mujtaba
//wccftech.com/intel-xeon-e52600-v3-haswellep-workstation-server-processors-unleashed-highperformance-computing

Intel Haswell-EP Xeon E5-2699 v3, die photo



HCC die which is reserved for the flagship parts including the Xeon E5-2699 V3 ranges from 14-18 core variants and configured on a 662 mm^2 consisting of a mind boggling 5.69 Billion transistors (English billion = 1 G = 10^9)

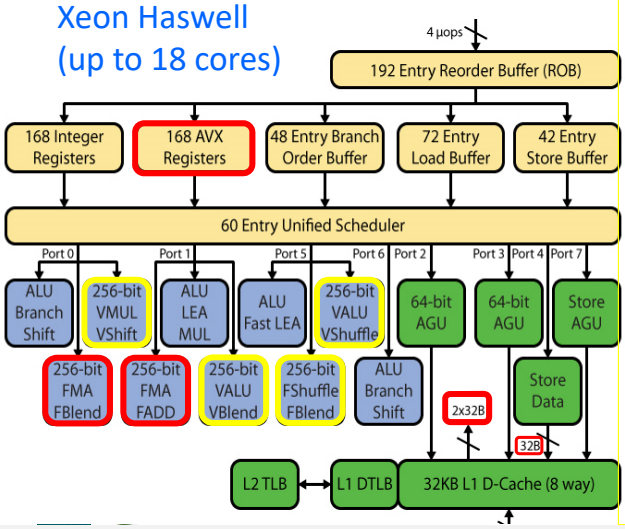
22 nm, 0.65V–1.30V, 2.3-3.6 GHz

TDP 145 W

Up to 4 DDR4 1600/1866/2133 Memory Channels

AVX is fully integrated in the execution core

Xeon Haswell
(up to 18 cores)



AVX-2, 256b → VL=32B

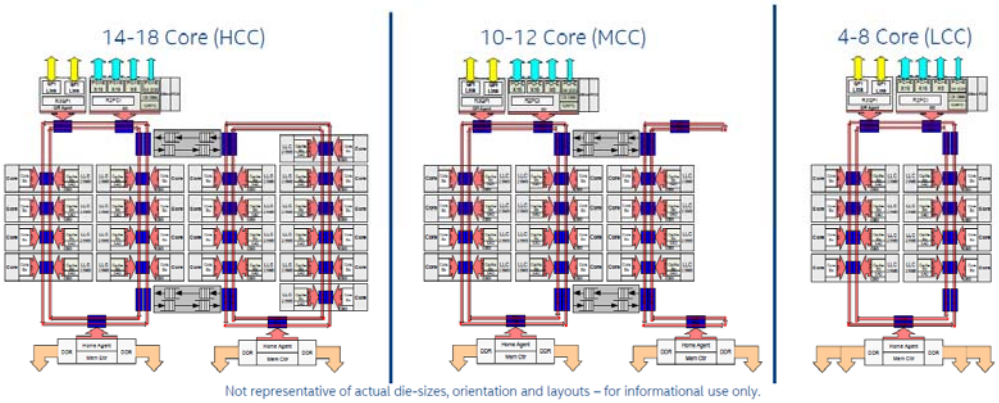
- Saves dynamic energy (many scalar overhead instructions removed)
- Saves static energy (computation ends early)
- 6 Vector Units:
  - 4 integers
  - 2 floats
- If a program …
  - reuses data in vector regs.
  - performs sequential memory traversal
  - …

  R may reach:
  16 DP FLOP/cycle
  32 SP FLOP/cycle

Intel's Haswell CPU Microarchitecture, November 13, 2012 by David Kanter, www.realworldtech.com/haswell-cpu/
Intel Xeon E5-2600 v3 (Haswell) Architecture &Features, 2014, Rama Karedla

14

Haswell EP Die Configurations



Not representative of actual die-sizes, orientation and layouts – for informational use only.

Chop	Columns	Home Agents	Cores	Power (W)	Transitors (B)	Die Area (mm²)
HCC	4	2	14-18	110-145	5.69	662
MCC	3	2	6-12	65-160	3.84	492
LCC	2	1	4-8	55-140	2.60	354

intel | 12

CANTABRIA CAMPUS INTERNACIONAL

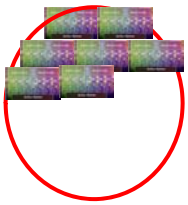
15
04/12/2017

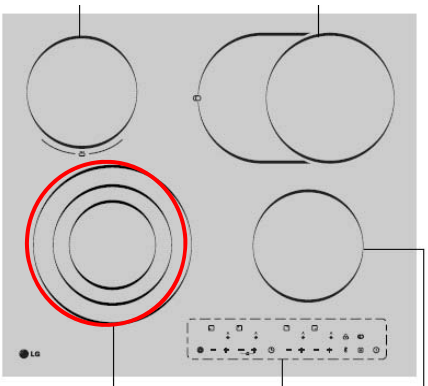
"Cocinando con el Xeon E5-2699 v3"

Vitro LG KA60530A: **2,3kW** en 3,14 x 110^2 = 38013,36 mm^2

Xeon E5-2699 v3: 145W en 662 mm^2

- caben unos ……
- disiparía ……..
- prestaciones en FLOP/ciclo ?





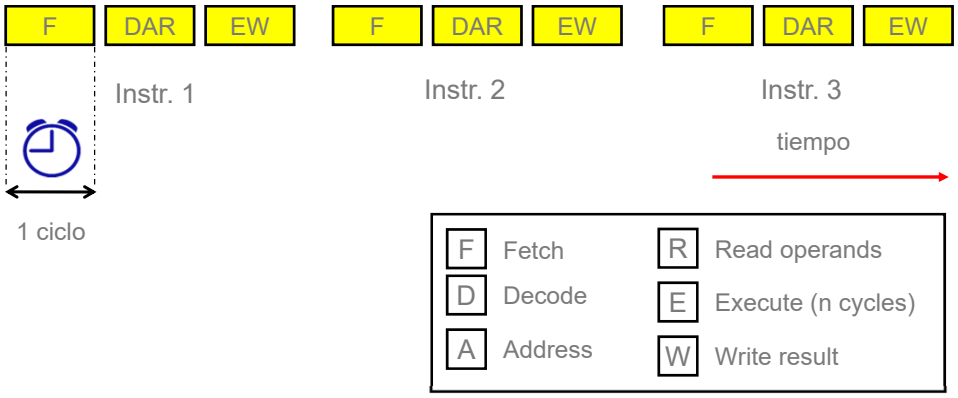
CANTABRIA CAMPUS INTERNACIONAL

Outline

- 1. Programs, Performance & Power
- 2. ISA Execution Models
- 3. Memory Basis
- 4. On-chip networks (NoCs)

Evolución en el modelo de ejecución

- Modelo 1 (3 **c**iclos **p**or **i**nstrucción, CPI=3),
  1ª Generación computadores, ejecución sin solape



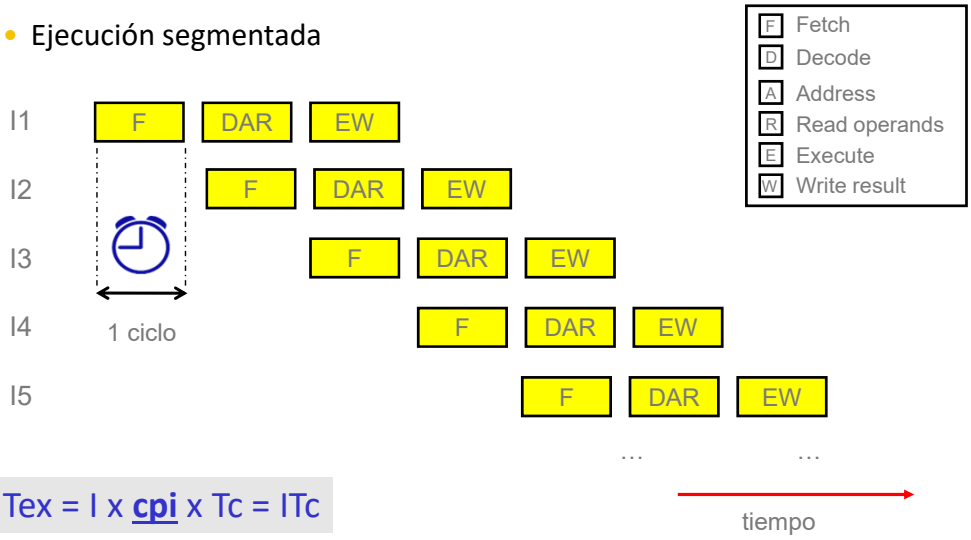
Tex = I x cpi x Tc=3ITc

Fases de ejecución de una instrucción (ejemplo)

Evolución en el modelo de ejecución

2 de 6

- Modelo 2 (CPI>=1)
- Ejecución segmentada

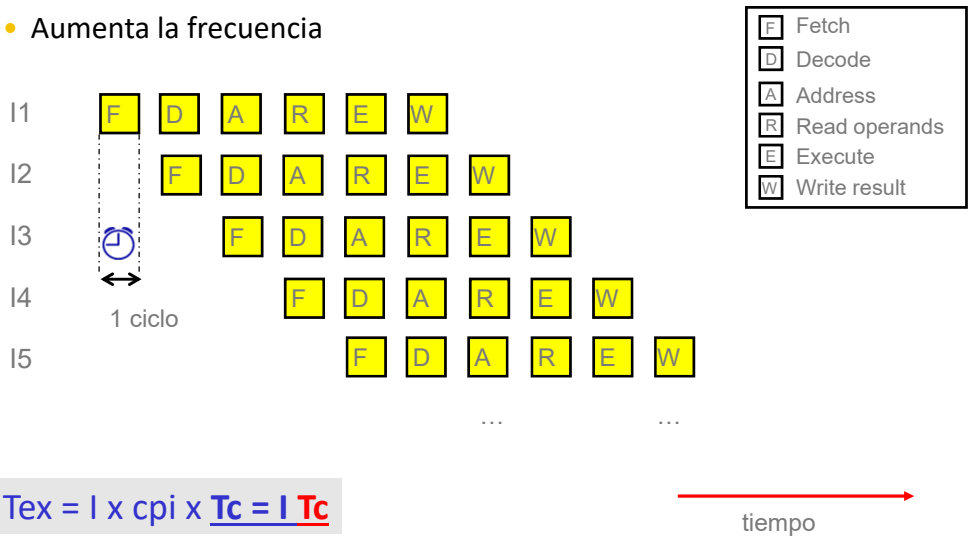


Tex = I x **cpi** x Tc = ITc

Evolución en el modelo de ejecución

3 de 6

- Modelo 3 (CPI>=1 pero Tc menor)
- Aumenta la frecuencia

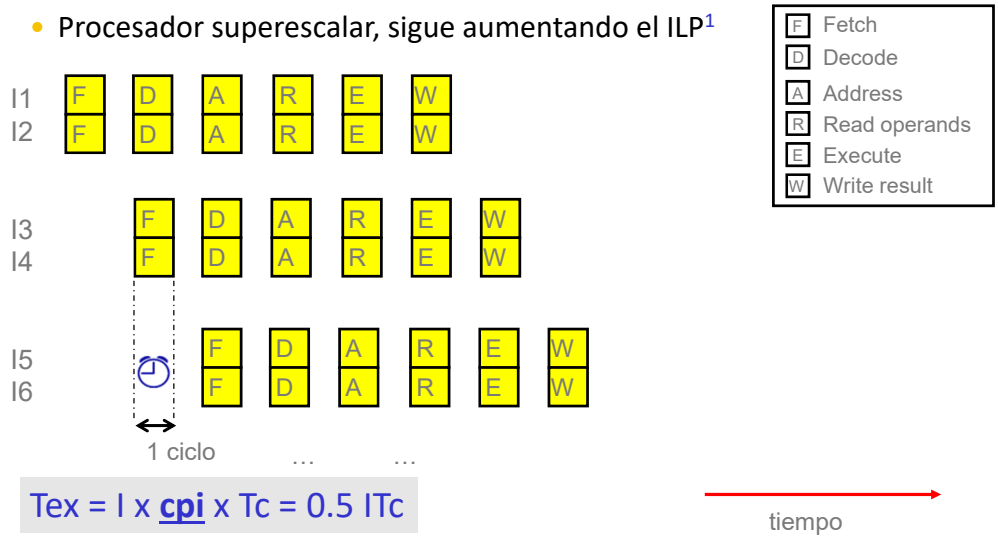


Tex = I x cpi x **Tc = I Tc**

Evolución en el modelo de ejecución

4 de 6

- Modelo 4 (CPI >=1/2, Tc pequeño) – en µ a finales de los 80 –
- Procesador superescalar, sigue aumentando el ILP[1]



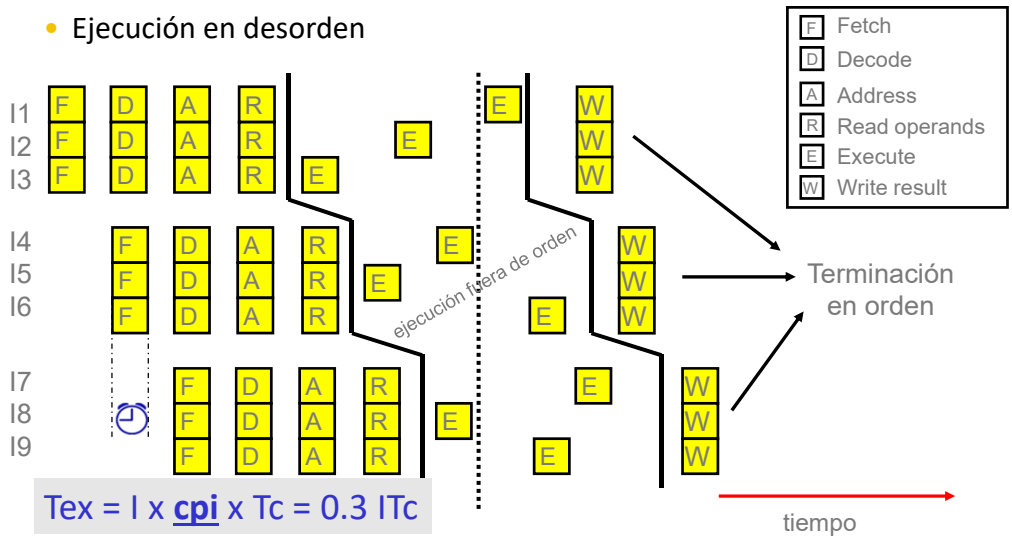
Tex = I x **cpi** x Tc = 0.5 ITc

1. ILP = Instruction Level Parallelism

Evolución en el modelo de ejecución

5 de 6

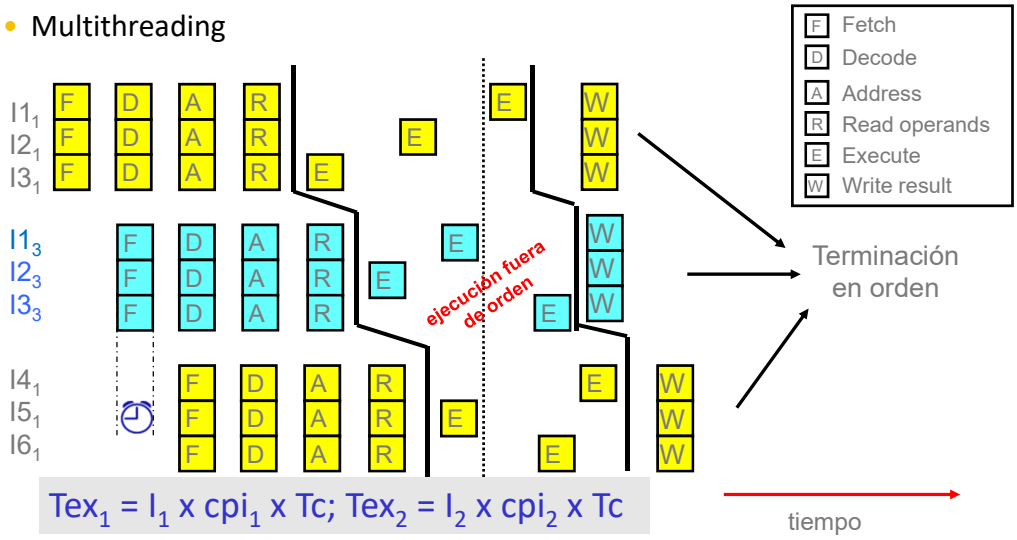
- Modelo 5 (≈ 0.3 cpi) – en µ a principios de los 90 –
- Ejecución en desorden



Tex = I x **cpi** x Tc = 0.3 ITc

Evolución en el modelo de ejecución

- Modelo 6 (>=0.16 cpi) – en µ en los años 2000 –
- Multithreading



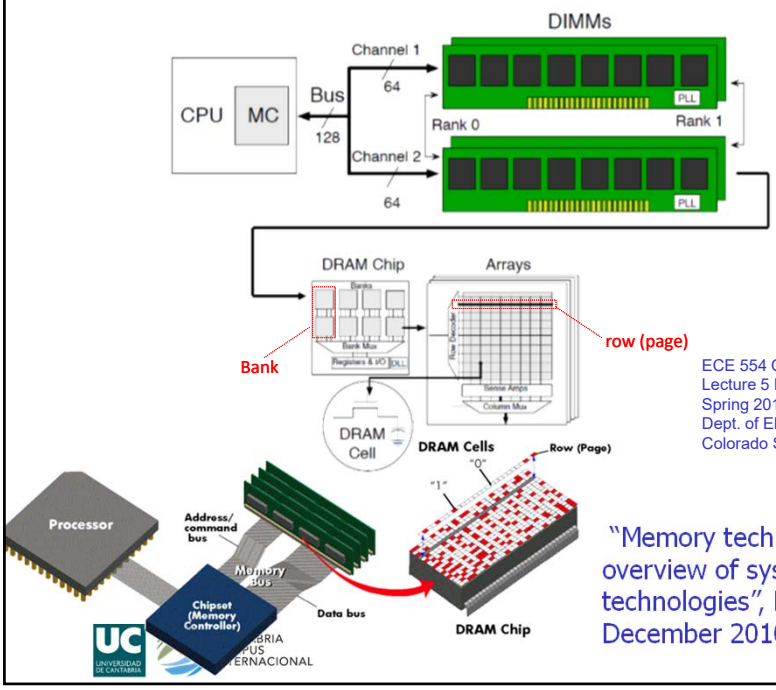
$Tex_1 = I_1 \times cpi_1 \times Tc$; $Tex_2 = I_2 \times cpi_2 \times Tc$

CACHES & BRANCH PREDICTORS CRITICAL!!!

Outline

1. Programs, Performance & Power
2. ISA Execution Models
3. Memory Basis
4. On-chip networks (NoCs)

Organización de la memoria principal



Controller
- Channel
- DIMM
- Rank
- Bank
- Page

!!

ECE 554 Computer Architecture.
Lecture 5 Main Memory.
Spring 2013. Sudeep Pasricha.
Dept. of Electrical and Computer Engineering
Colorado State University

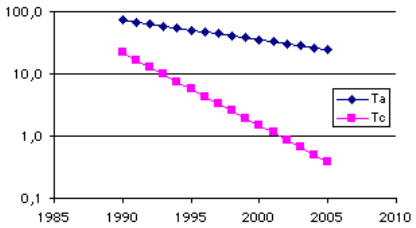
"Memory technology evolution: an overview of system memory technologies", HP Technology brief, December 2010.

25

Discrepancia entre procesador y memoria DRAM

Speed gap,

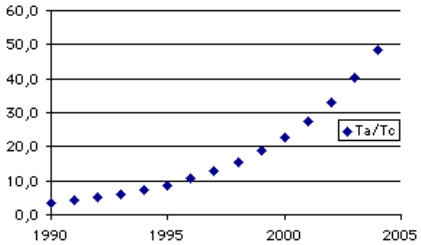
memory wall



- Tiempo de acceso a memoria medido en ciclos de procesador:

$$\frac{\text{Tiempo de acceso a memoria}}{\text{Tiempo de ciclo de procesador}}$$

+21% cada año



CANTABRIA CAMPUS INTERNACIONAL

26

Memoria SRAM vs. DRAM

	Celda de almacenamiento	Capacidad (chip 2008)	Precio/MByte	Velocidad
DRAM	1 Condensador + 1 Transistor alta densidad	grande 2048 Mbits	barato 0,18$	lenta 12.5 nsg
SRAM	4Trans. + 2Trans. baja densidad (16-25x DRAM)	pequeña 32 Mbits	caro ~ 18$	rápida 0.4 nsg

- Ejemplo: 2048 Mbits de memoria principal para un PC
  - DRAM: 16 chips, 20 €
  - SRAM: 64 chips, ~2.000 € (¡ sólo los chips !)
- Otros problemas de SRAM:
  - Encaminamiento aumentaría mucho el tiempo de acceso de SRAM
  - Consumo mucho mayor en SRAM

The Stacked Capacitor DRAM Cell and Three-Dimensional Memory. Mitsumasa Koyanagi, January 2008 IEEE Solid State Circuits Society Journal.

Caches

Propiedades de las memorias
- DRAM barata y lenta
- SRAM cara y rápida

Propiedad de los programas
- Localidad espacial y temporal
- regla 90/10 instrucciones

Solución: dos almacenes
- Mem. principal: DRAM, grande y lenta

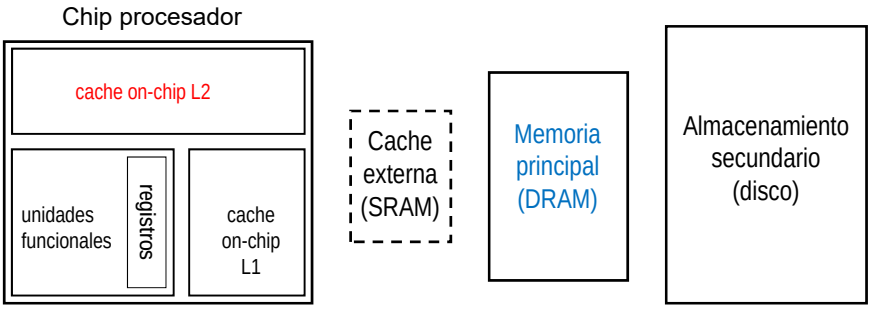
  almacén global
- Mem. cache: SRAM, pequeña y rápida

  subconjunto más usado

Memoria Cache: memoria pequeña y rápida que contiene el subconjunto más usado de memoria principal

Jerarquía de memoria

Varios niveles de almacenamiento



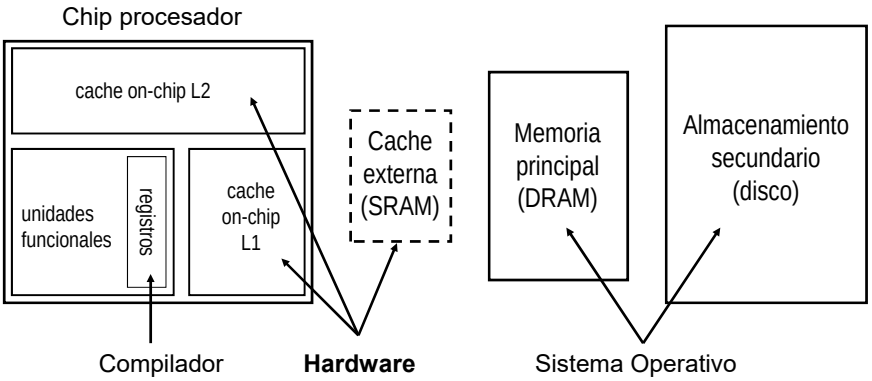
Tiempo (ciclos)	1	1-2	5-10	40	100-200	10 M
Capacidad (bytes)	160 * 8	8-64 KB	~ 1 MB	~ 32 MB	Gigabytes	Terabytes

incremento de velocidad ←

incremento de capacidad →

Jerarquía de memoria

Responsable de cada nivel

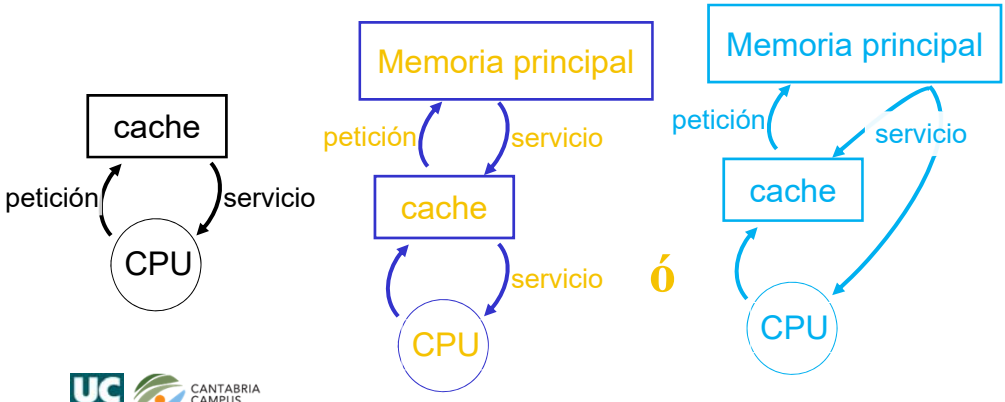


La memoria cache es **transparente** al nivel de lenguaje máquina

Descripción funcional

Acceso del procesador a CADA palabra (INSTRUCCIONES Y DATOS SEPARADOS)

- Primero se mira en cache
  - Acierto (hit) : se encuentra la palabra buscada (CPU + x' ó Mc + x')
  - Fallo (miss): no se encuentra la palabra buscada
- Sólo cuando hay MISS se accede a memoria principal



Descripción funcional: gestión de contenidos

Capacidad de Cache << capacidad de DRAM

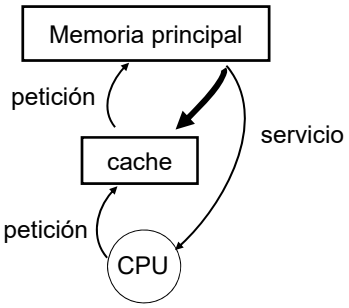
- ¿qué datos se guardan en cache?

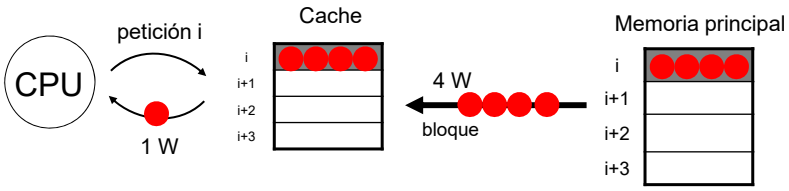
Aprovechar localidad temporal

- Si CPU pide una palabra, probablemente la volverá a pedir
  - Guardar palabra en cache

Aprovechar localidad espacial

- Si CPU pide una palabra, probablemente pedirá las cercanas
  - Guardar varias palabras contiguas: bloque





Prestaciones

Tasas

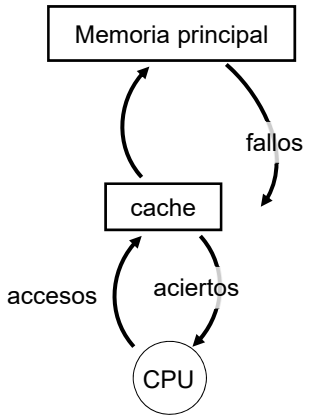
- **Tasa de fallos (m)**

  m = fallos / accesos

- **Tasa de aciertos (h)**

  h = aciertos / accesos

  h + m = 1



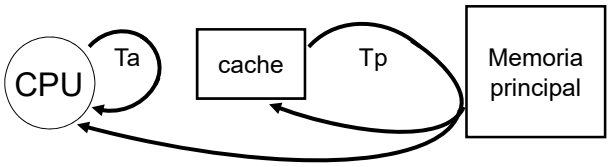
Costes (en tiempo o en nº de ciclos)

Coste de acierto: Ta (ns) o Ca (ciclos)

- Mirar en cache y entregar palabra al procesador

Penalización de fallo: Tp (ns) o Cp (ciclos)

- Desde que se detecta un fallo hasta que se sirve la palabra al procesador (leer bloque de MP, escribir bloque en cache …)



Coste efectivo de acceso a memoria (Tef, Cef)

Tiempo medio de acceso a memoria

Acierto — Ta

Fallo — Ta | Tp

$$Tef = h*Ta + m(Ta + Tp)$$

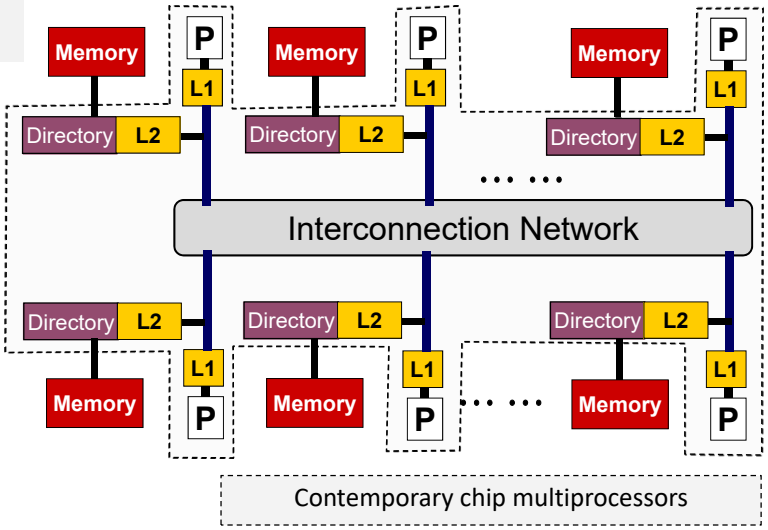
$$Tef = Ta + \underline{m*Tp}$$ penalización media

- Ejemplo

h = 0.9 m = 0.1	Sin cache	Cef = 50 ciclos
Ca = 1 ciclo Cp = 50 ciclos	Con cache	Cef = 1 + 0.1*50 = 6 ciclos

Current MP-SoCs

- 1 directory entry per shared last-level cache block



L1 Privadas

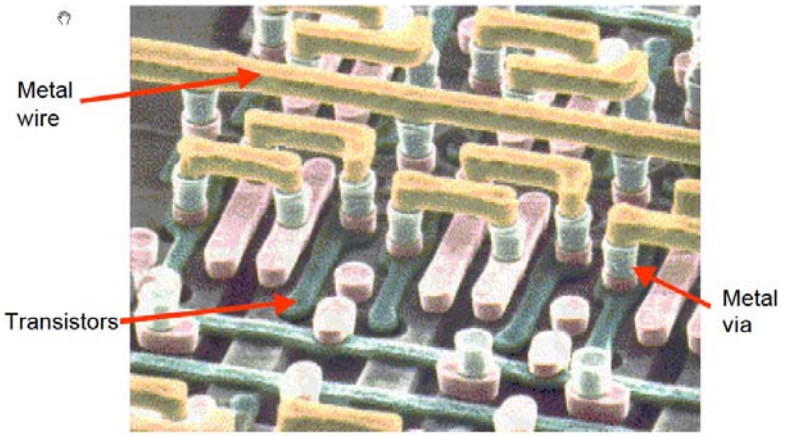
L2 Compartidas "Homes"

Outline

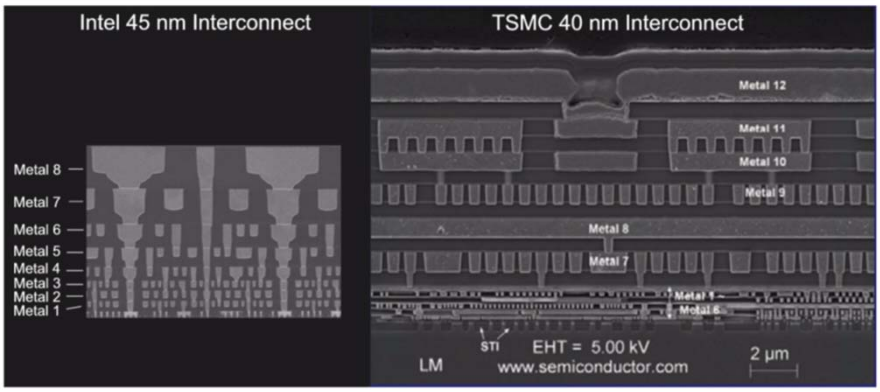
- 1. Programs, Performance & Power
- 2. ISA Execution Models
- 3. Memory Basis
- 4. On-chip networks (NoCs)

On-Chip local interconnects

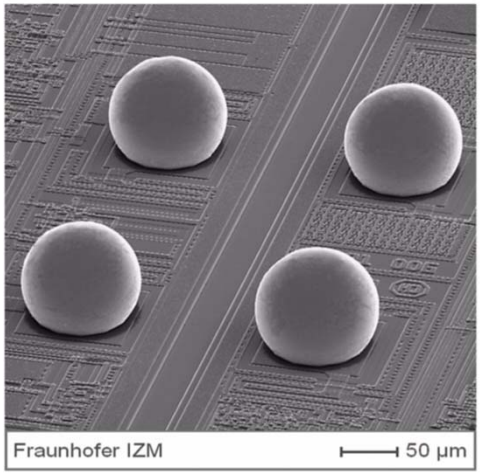
SEM photo of local levels interconnect

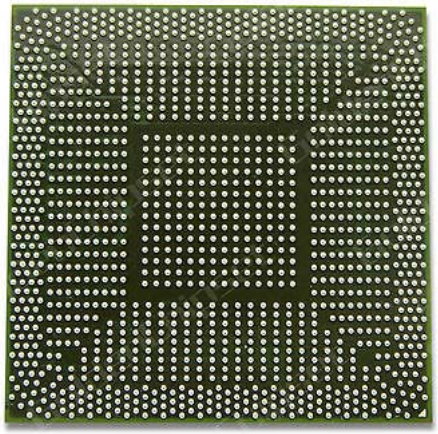


Metal Layers



Bumps & Balls

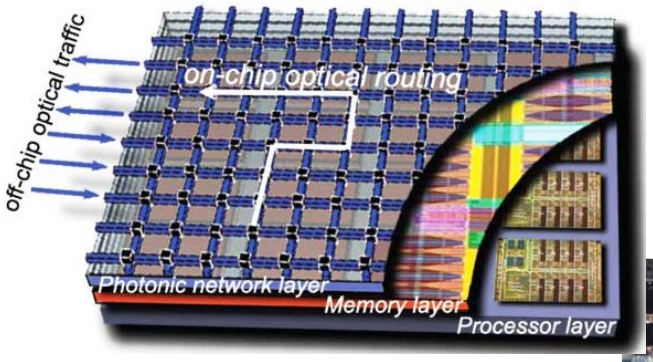


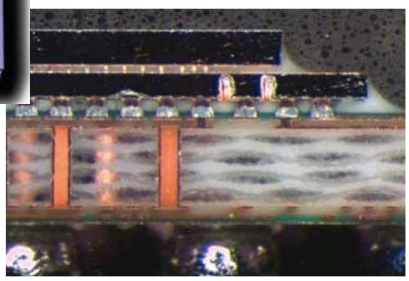


goldenchipset

3D (& 2.5D) Stacking & Silicon Photonics

Multiple integration with 3D stacking…



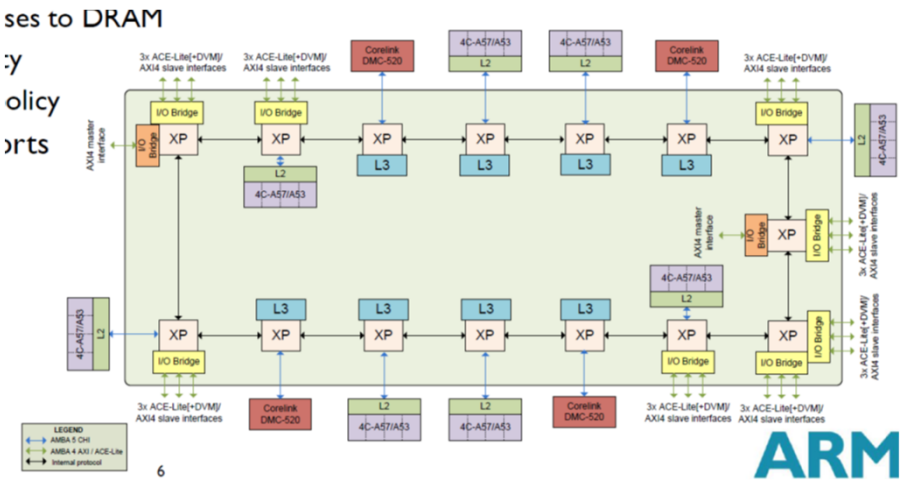


3M, IBM team to develop 3D IC adhesive, EETimes India

STMicroelectronics & CEA

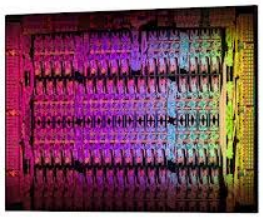
CANTABRIA CAMPUS INTERNACIONAL

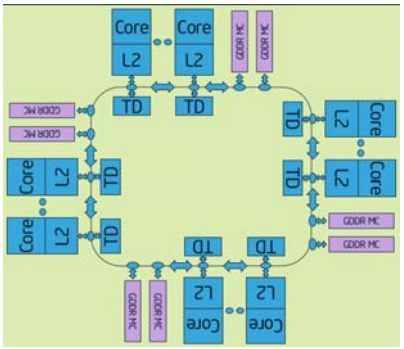
Rings from ARM



CANTABRIA CAMPUS INTERNACIONAL

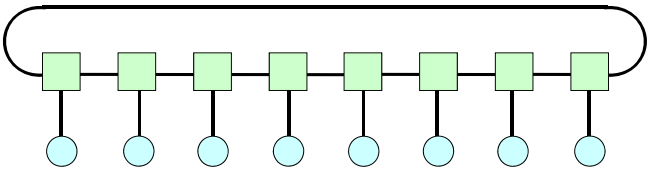
Rings from Intel

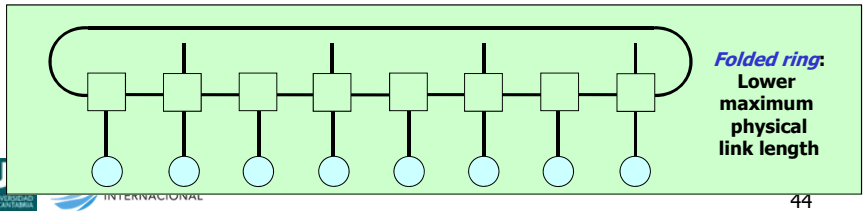




Rings (Direct or Indirect?)

- ***Bidirectional Ring networks (folded)***
  - *N* switches (3 × 3) and *N* bidirectional network links
  - Simultaneous packet transport over disjoint paths
  - Packets must hop across intermediate nodes
  - Shortest direction usually selected (*N*/4 hops, on average)
  - Bisection Bandwidth???

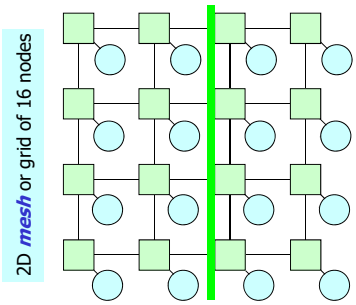


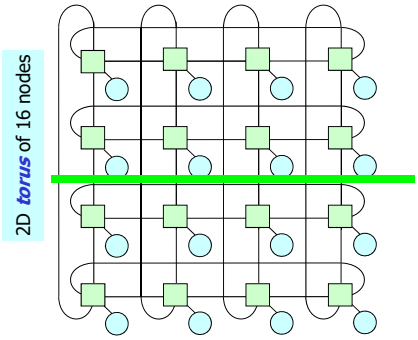


***Folded ring*:** **Lower maximum physical link length**

Meshes and Tori

Distributed Switched (Direct) Networks



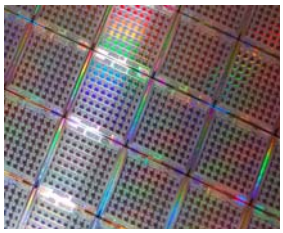


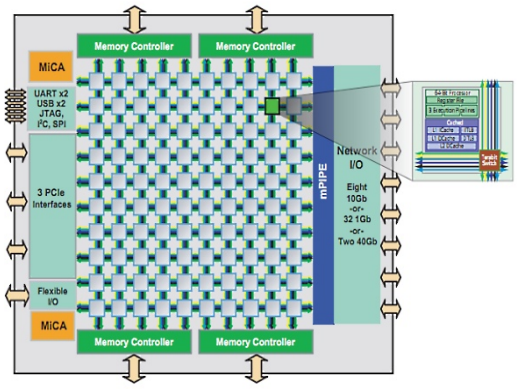
Network Bisection

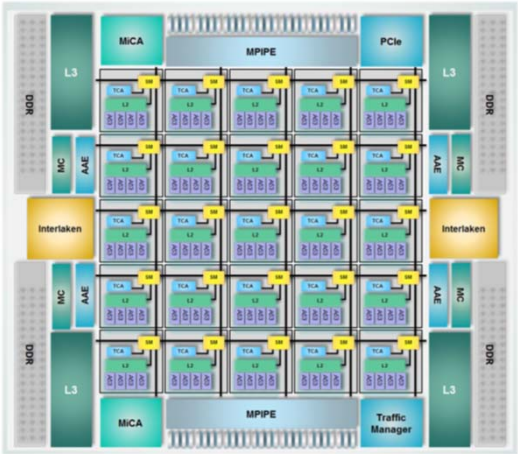
CANTABRIA CAMPUS INTERNACIONAL

45

Meshes from Tilera



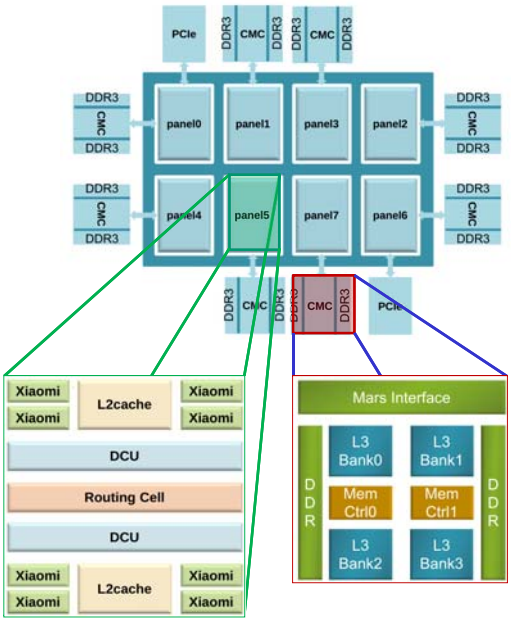




CANTABRIA CAMPUS INTERNACIONAL

Mesh from Pythium Mars Architecture

- L1:
  - Separated L1 Icache and L1 Dcache
  - 32 KB Icache
  - 32 KB Dcache
    - 6 outstanding loads
    - 4 cycles latency from load to use
- L2:
  - 16 L2 banks of 4 MB
  - 32 MB of shared L2
- L3:
  - 8 L3 arrays of 16 MB
  - 128 MB of L3
- Memory Controllers:
  - 16 DDR3-1600 channels
- 2x16-lane PCIe-3.0
- Directory based cache coherency
  - 16 Directory Control Unit (DCU)
- MOESI like cache coherence protocol



These images were taken form the slides presented at Hot Chips 2015

Meshes from Intel Knights Landing

Knights Landing: Next Intel® Xeon Phi™ Processor

Intel® Many-Core Processor targeted for HPC and Supercomputing

First **self-boot** Intel® Xeon Phi™ processor that is **binary compatible** with main line IA. Boots standard OS.

Significant improvement in scalar and **vector** performance

Integration of **Memory on package**: innovative memory architecture for high bandwidth and high capacity

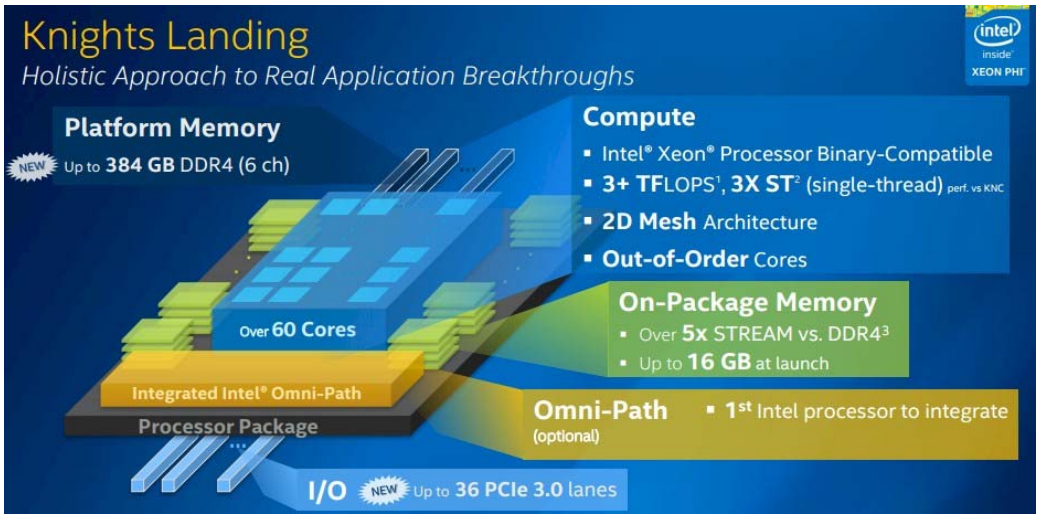
Integration of **Fabric on package**

Three products

KNL Self-Boot	KNL Self-Boot w/ Fabric	KNL Card
(Baseline)	(Fabric Integrated)	(PCIe-Card)

Potential future options subject to change without notice.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notification.

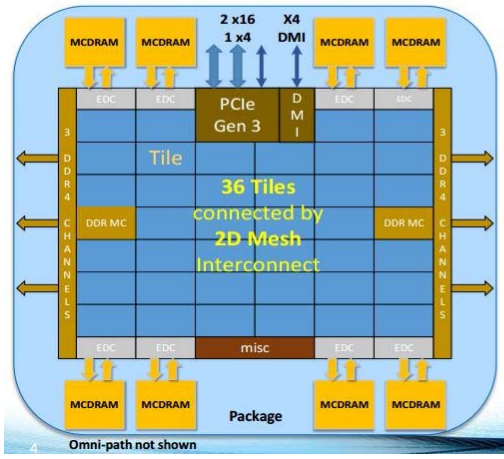
Intel Knights Landing



CANTABRIA CAMPUS INTERNACIONAL

Intel Knights Landing

Knights Landing Overview



Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384GB
IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset
Node: 1-Socket only
Fabric: Omni-Path on-package (not shown)

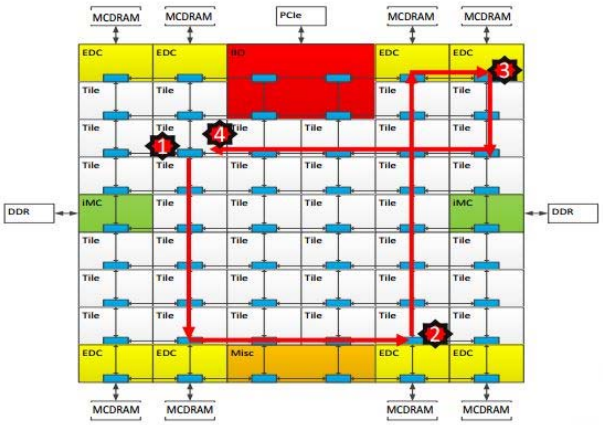
Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set Architecture (ISA). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

CANTABRIA CAMPUS INTERNACIONAL

Intel Knights Landing

Cluster Mode: All-to-All



Address uniformly hashed across all distributed directories

No affinity between Tile, Directory and Memory

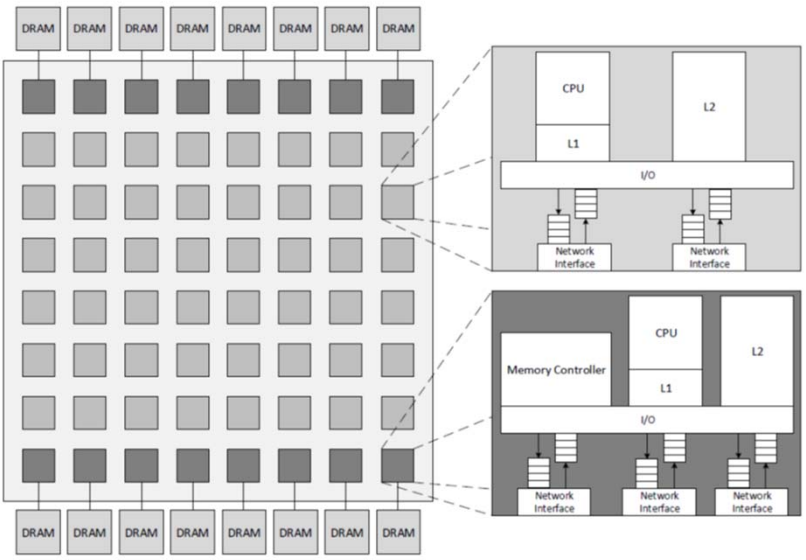
Most general mode. Lower performance than other modes.

Typical Read L2 miss

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requestor

CANTABRIA CAMPUS INTERNACIONAL

NUCA-based CMP Models



CANTABRIA CAMPUS INTERNACIONAL

52