



Master Informatics Eng.

2018/19

A.J.Proença

Concepts from undergrad Computer Systems (1)

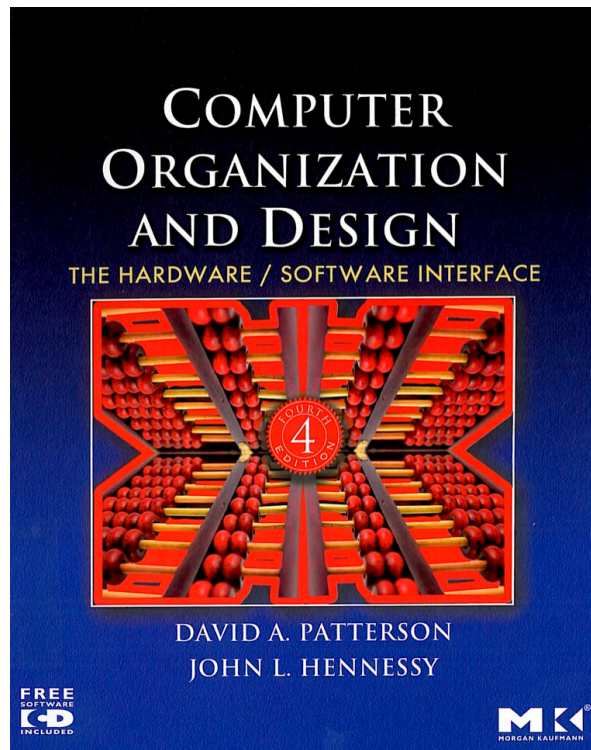
(most slides are borrowed, mod's in green)

Advanced Architectures



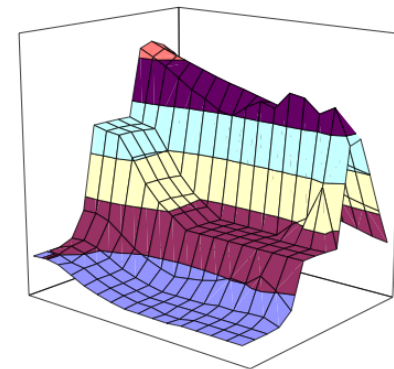
Concepts from undergrad Computer Systems

– *most slides are borrowed from*



and some from

Computer Systems
*A Programmer's Perspective*¹
(Beta Draft)



Randal E. Bryant
David R. O'Hallaron

August 1, 2001

more details at
<http://gec.di.uminho.pt/miei/sc/>

Background for Advanced Architectures



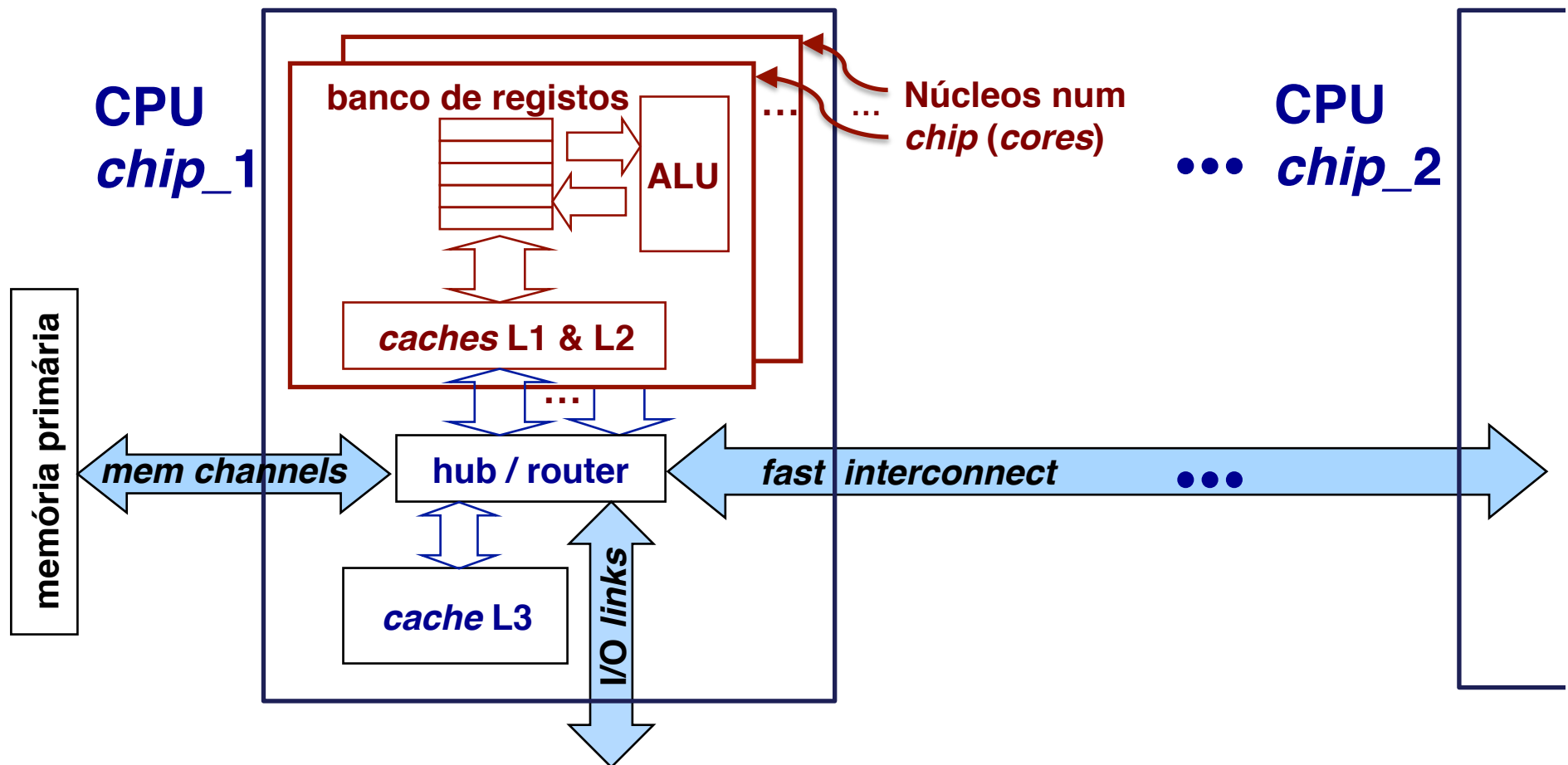
Key concepts to revise:

- *numerical data representation (for error analysis)*
- *ISA (Instruction Set Architecture)*
- *how C compilers generate code (a look into assembly code)*
 - *how scalar and structured data are allocated*
 - *how control structures are implemented*
 - *how to call/return from function/procedures*
 - *what architecture features impact performance*
- ***Improvements to enhance performance in a single CPU***
 - ***ILP: pipeline, multiple issue, ...***
 - ***data parallelism: SIMD/vector processing, ...***
 - ***memory hierarchy: cache levels, ...***
 - ***thread-level parallelism***

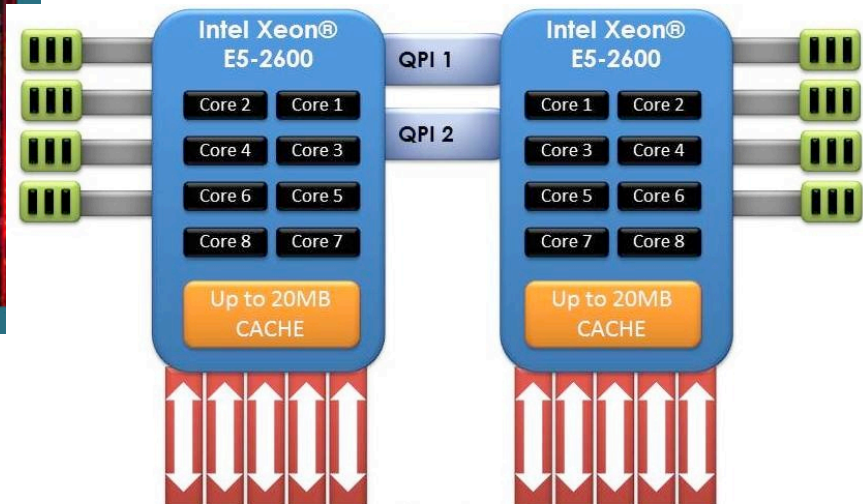
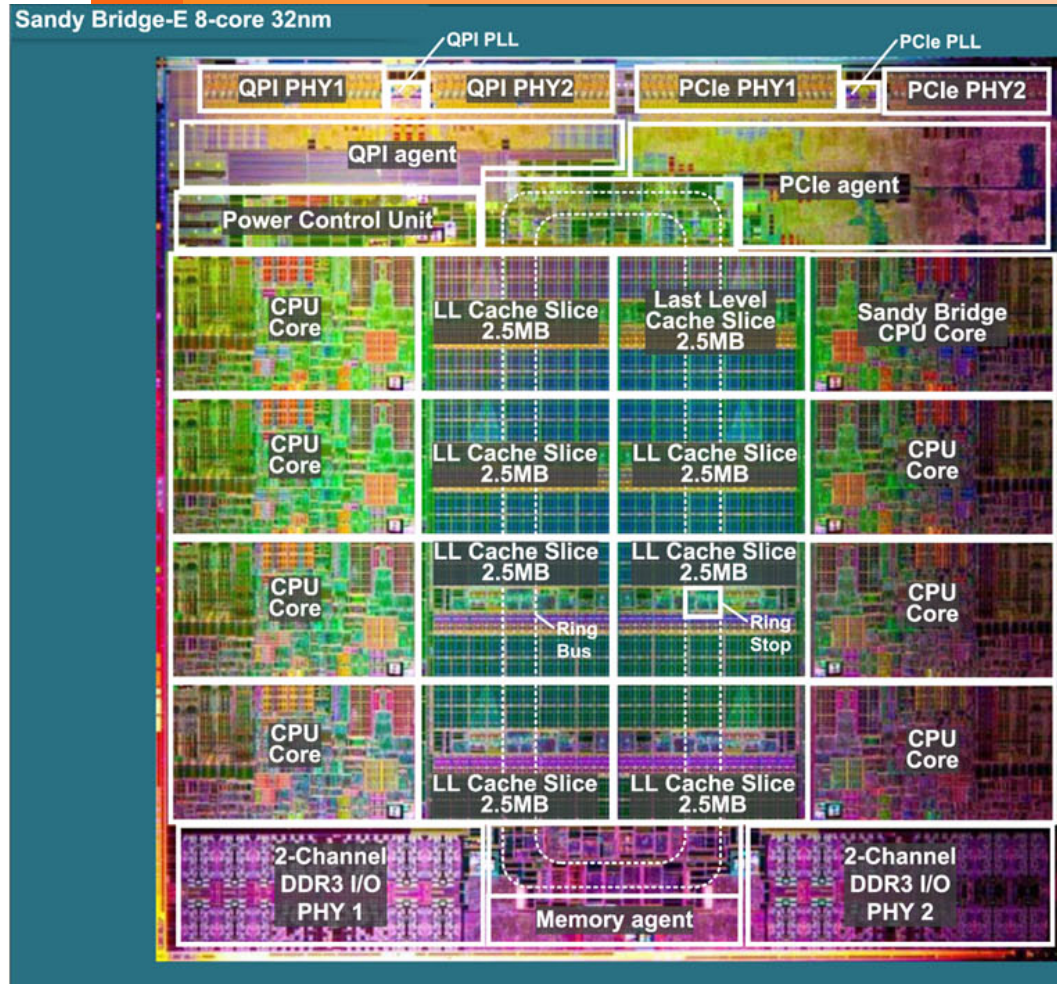
A hierarquia de cache em arquiteturas multicore



As arquiteturas *multicore* mais recentes:



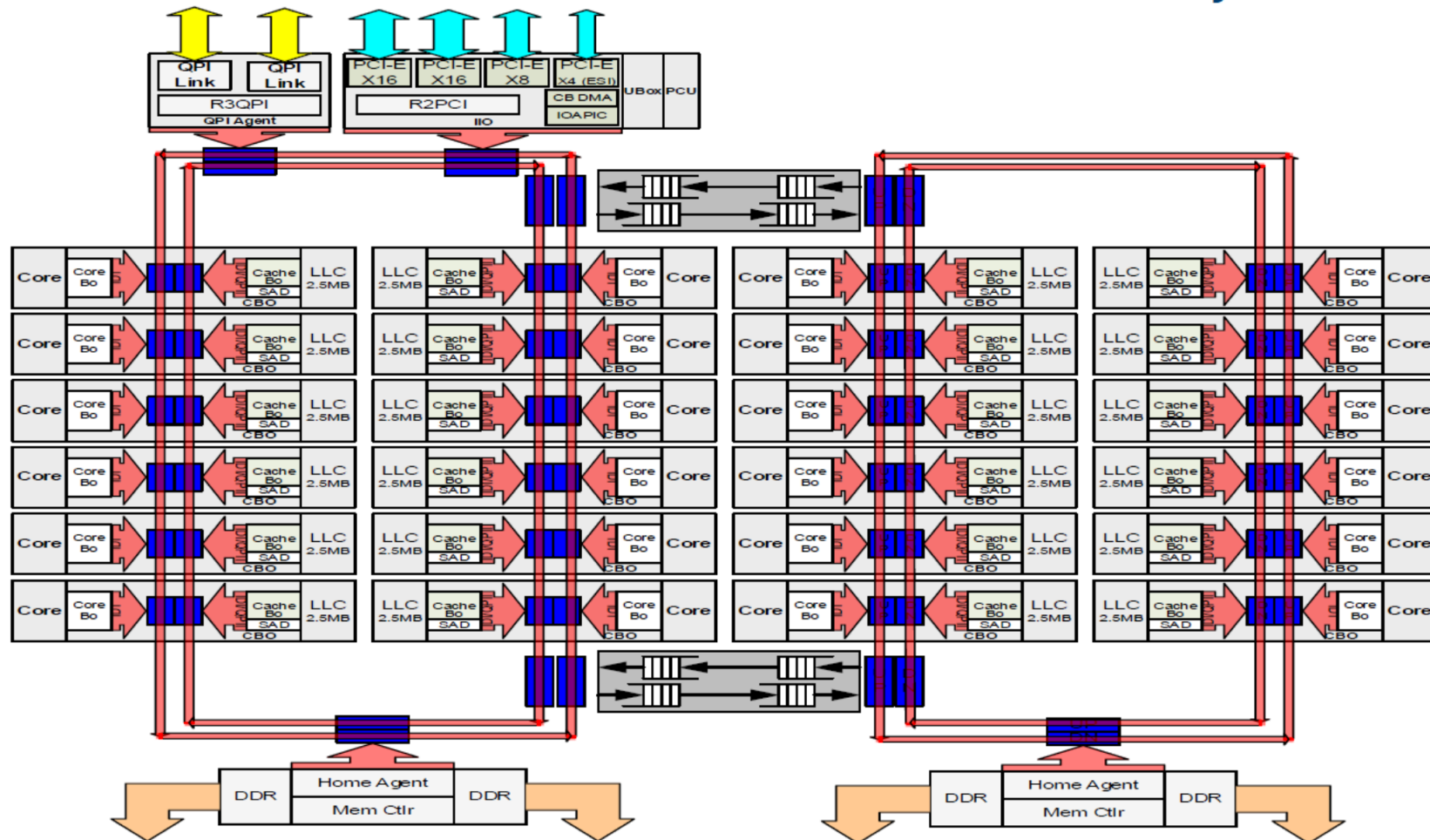
Lançamento da Intel em 2012: Sandy/Ivy Bridge (8-core)



Intel in 2016:
Broadwell-EP Xeon (22-core)

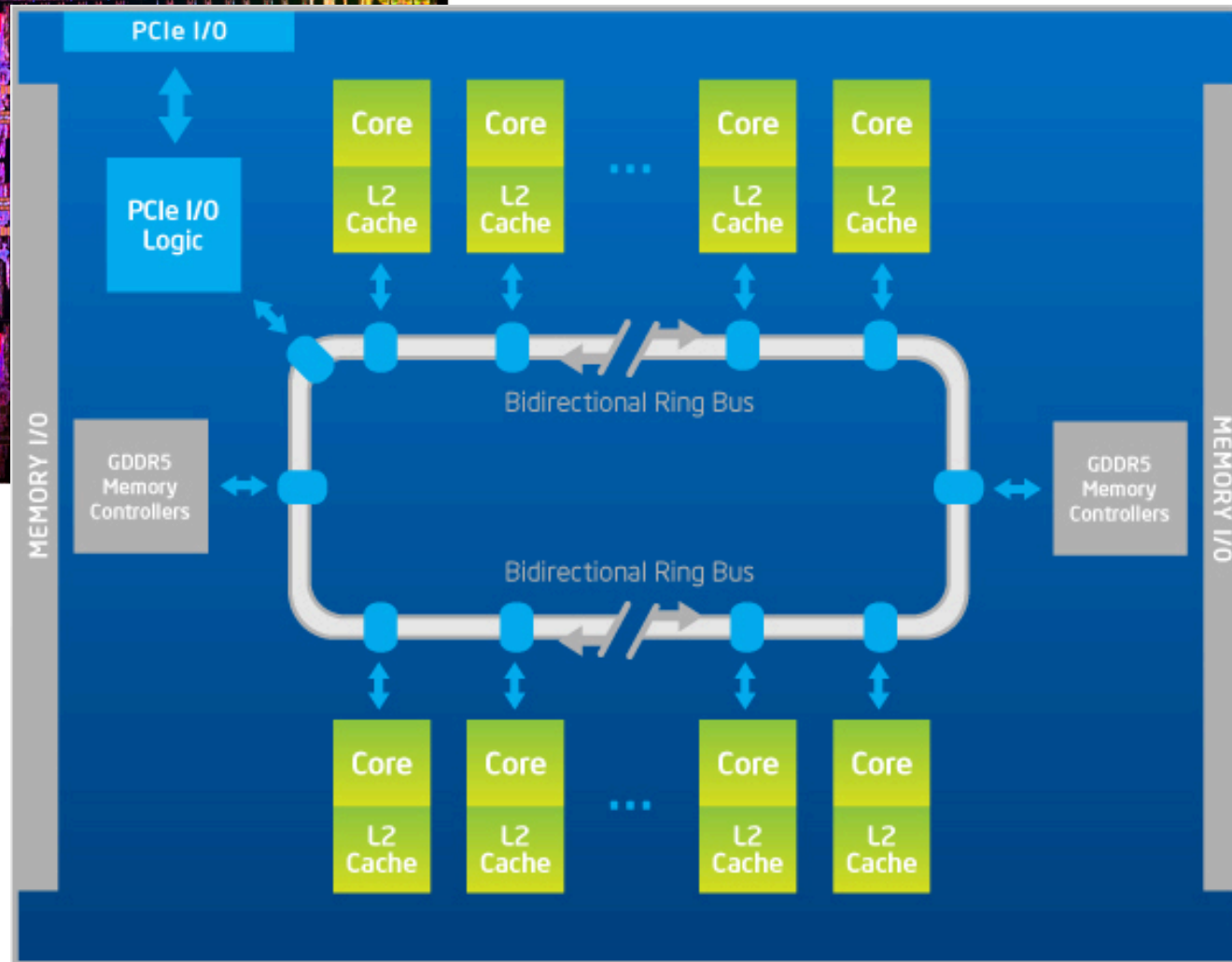


Intel® Xeon® Processor E5 v4 Product Family HCC





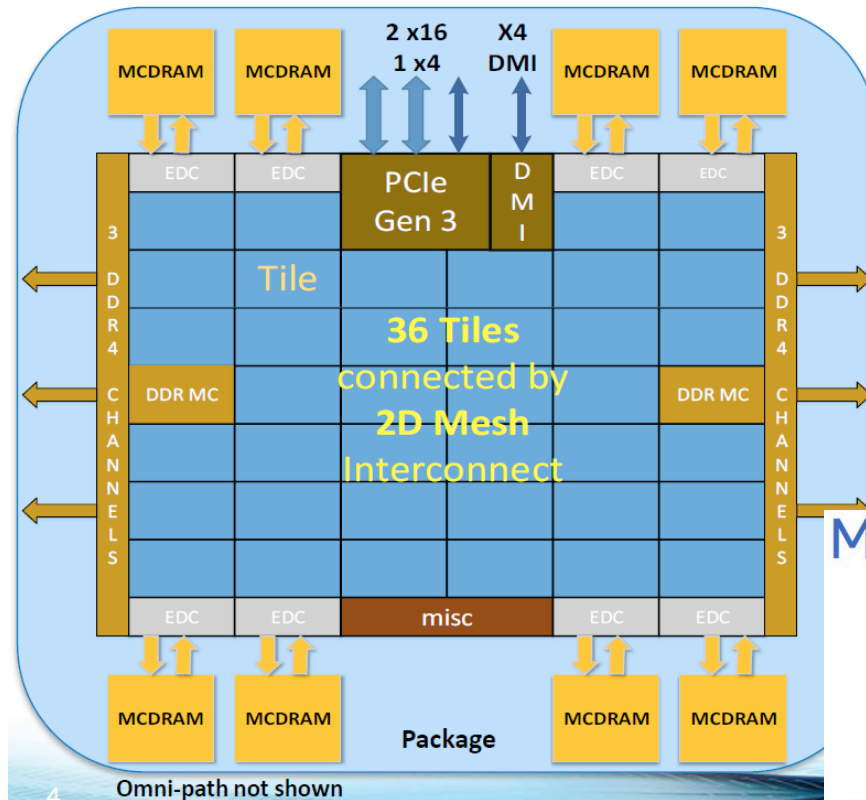
Chips da Intel em 2012/13: Xeon Phi com 60 cores



Intel new Phi in 2016: KNL with 72 cores



Knights Landing Overview



TILE

2 VPU	CHA	2 VPU
Core	1MB L2	Core

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

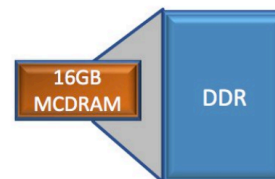
Node: 1-Socket only

Fabric: Omni-Path on-package (not shown)

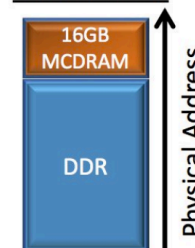
Memory Modes

Three Modes. Selected at boot

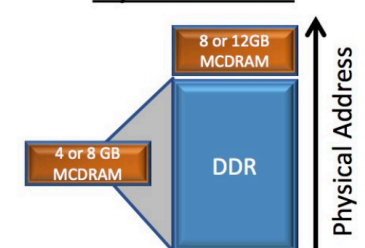
Cache Mode



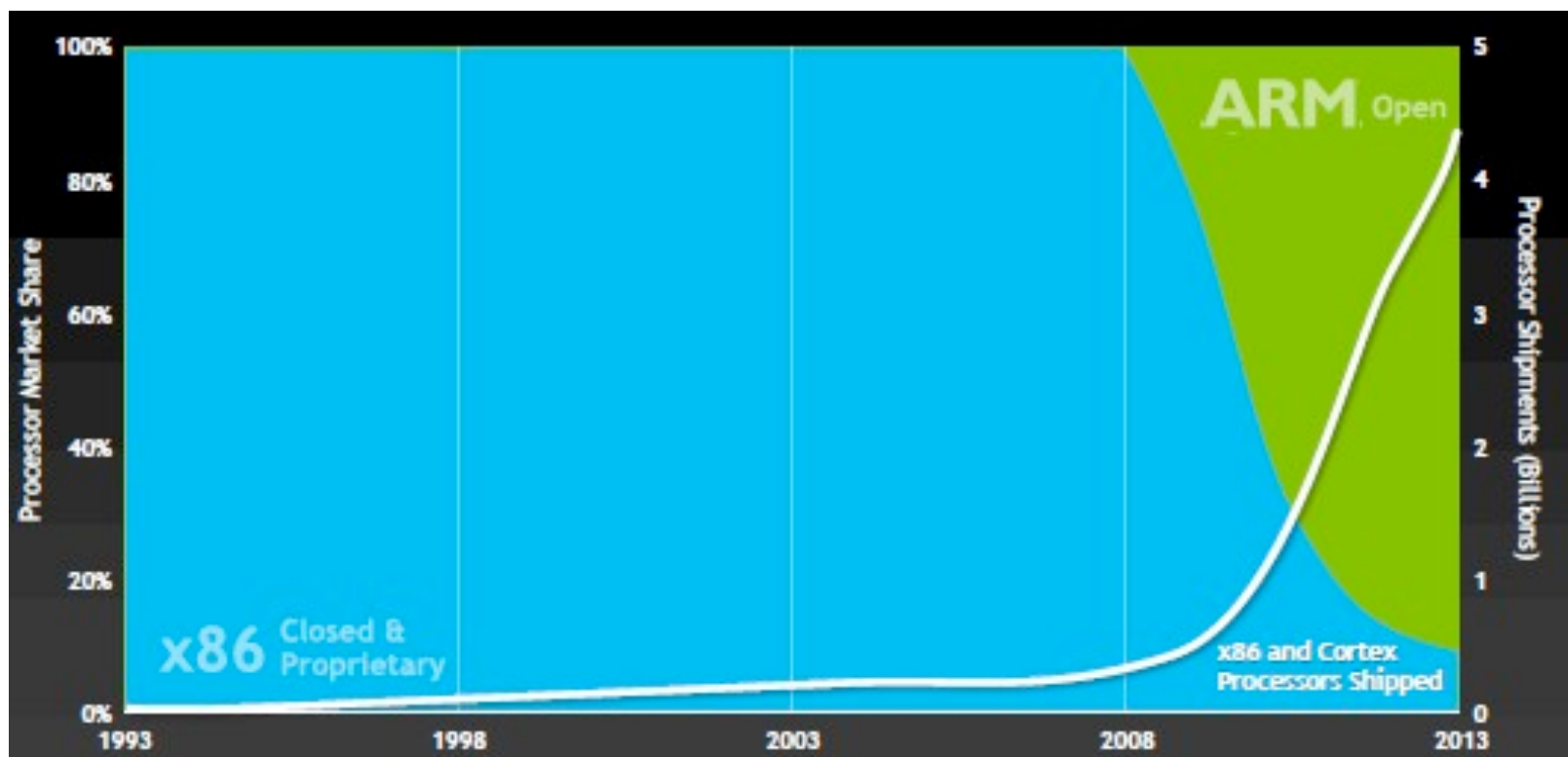
Flat Mode



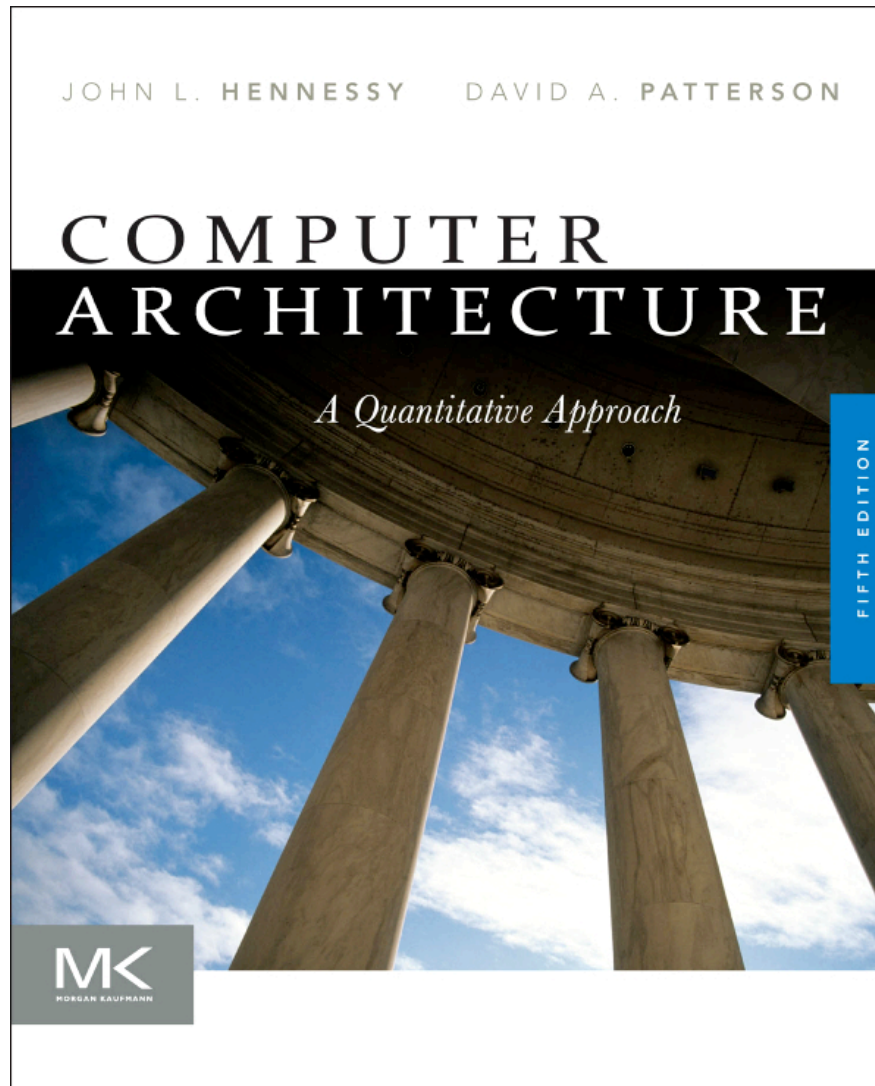
Hybrid Mode



Processadores Intel x86 versus ARM



Key textbook for AA



Computer Architecture, 5th Edition

Hennessy & Patterson

Table of Contents

Printed Text

- Chap 1: Fundamentals of Quantitative Design and Analysis
- Chap 2: Memory Hierarchy Design
- Chap 3: Instruction-Level Parallelism and Its Exploitation
- Chap 4: Data-Level Parallelism in Vector, SIMD, and GPU Architectures
- Chap 5: Multiprocessors and Thread-Level Parallelism
- Chap 6: The Warehouse-Scale Computer
- App A: Instruction Set Principles
- App B: Review of Memory Hierarchy
- App C: Pipelining: Basic and Intermediate Concepts

Online

- App D: Storage Systems
- App E: Embedded Systems
- App F: Interconnection Networks
- App G: Vector Processors
- App H: Hardware and Software for VLIW and EPIC
- App I: Large-Scale Multiprocessors and Scientific Applications
- App J: Computer Arithmetic
- App K: Survey of Instruction Set Architectures
- App L: Historical Perspectives

Recommended textbook (1)

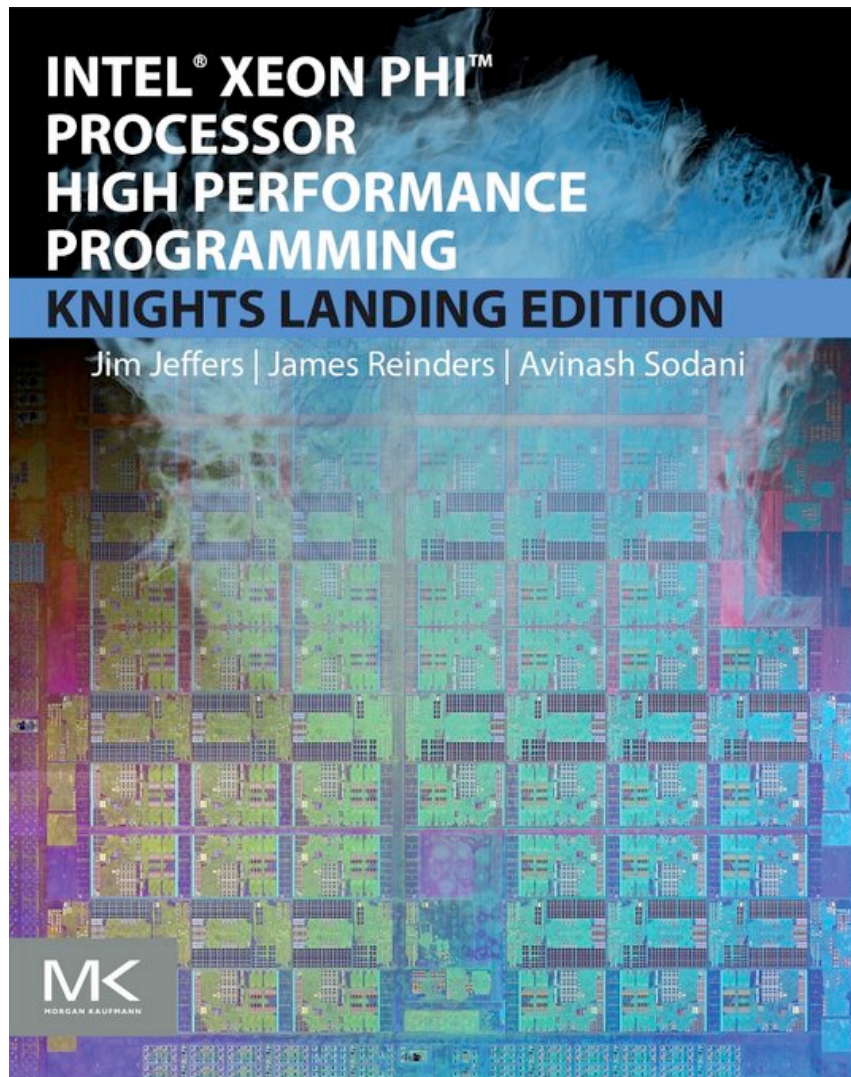


Table of Contents

Section I: Knights Landing.

- Chapter 1:** Introduction
- Chapter 2:** Knights Landing Overview
- Chapter 3:** Programming MCDRAM and Cluster Modes
- Chapter 4:** Knights Landing Architecture
- Chapter 5:** Intel Omni-Path Fabric
- Chapter 6:** March Optimization Advice

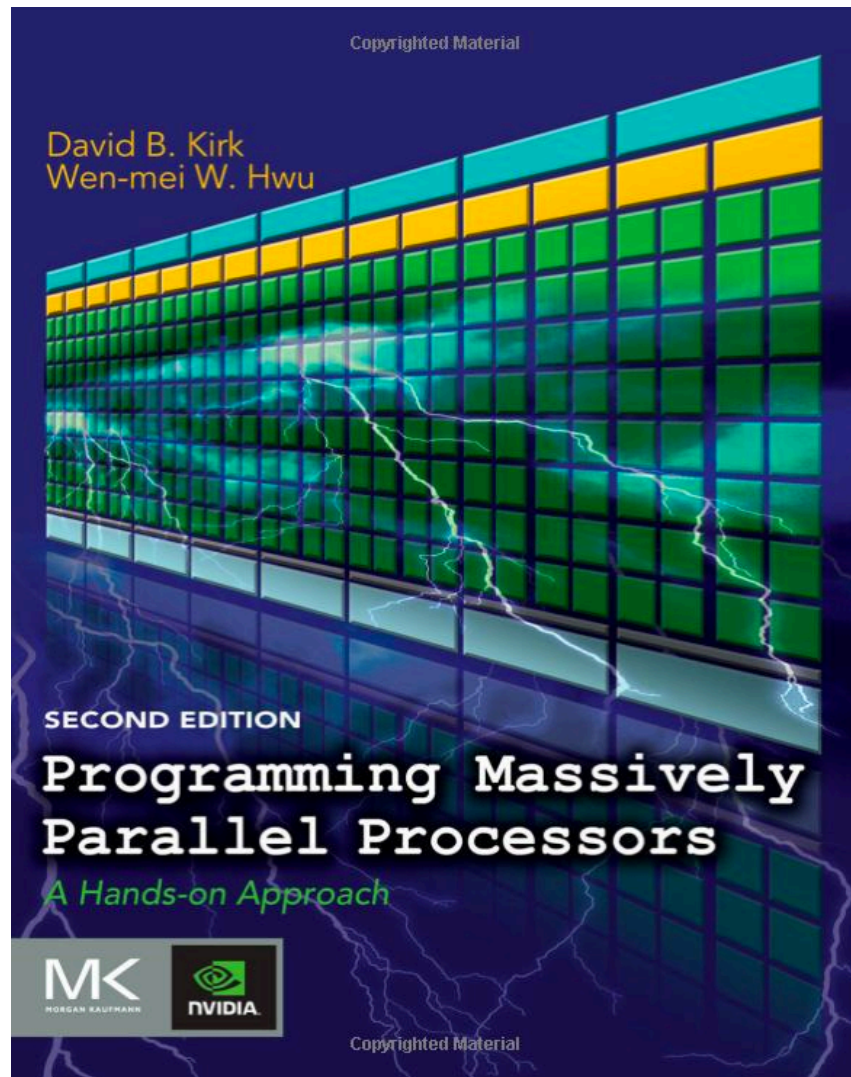
Section II: Parallel Programming

- Chapter 7:** Programming Overview for Knights Landing
- Chapter 8:** Tasks and Threads
- Chapter 9:** Vectorization
- Chapter 10:** Vectorization Advisor
- Chapter 11:** Vectorization with SDLT
- Chapter 12:** Vectorization with AVX-512 Intrinsics
- Chapter 13:** Performance Libraries
- Chapter 14:** Profiling and Timing
- Chapter 15:** MPI
- Chapter 16:** PGAS Programming Models
- Chapter 17:** Software Defined Visualization
- Chapter 18:** Offload to Knights Landing
- Chapter 19:** Power Analysis

Section III: Pearls

- Chapters 20-26:** Results on LAMMPS, SeisSol, WRF, N-Body Simulations, Machine Learning, Trinity mini-applications and QCD are discussed.

Recommended textbook (2)



Contents

- 1 Introduction
- 2 History of GPU Computing
- 3 Introduction to Data Parallelism and CUDA C
- 4 Data-Parallel Execution Model
- 5 CUDA Memories
- 6 Performance Considerations
- 7 Floating-Point Considerations
- 8 Parallel Patterns: Convolution
- 9 Parallel Patterns: Prefix Sum
- 10 Parallel Patterns: Sparse Matrix-Vector Multiplication
- 11 Application Case Study: Advanced MRI Reconstruction
- 12 Application Case Study: Molecular Visualization and Analysis
- 13 Parallel Programming and Computational Thinking
- 14 An Introduction to OpenCL
- 15 Parallel Programming with OpenACC
- 16 Thrust: A Productivity-Oriented Library for CUDA
- 17 CUDA FORTRAN
- 18 An Introduction to C11 AMP
- 19 Programming a Heterogeneous Computing Cluster
- 20 CUDA Dynamic Parallelism
- 21 Conclusion and Future Outlook