Apresentação Aulas Presenciais

Performance em Sistemas Ciberfísicos

Frank Coelho de Alcantara – 2023 -1

Apresentação da Disciplina

No Canvas

Um Pouco de História e Contexto

Aqueles que não conhecem a história
estão condenados a repeti-la.
Edmund Burke

Porque se Preocupar com o Hardware

- ***Entender o destino da computação:***
  - *O que podemos esperar da tecnologia;*
  - *Impacto da tecnologina no mundo real.*
- ***Entender os conceitos de design de mais alto nível:***
  - *Os melhores projetistas entendem todos os níveis.*
- ***Entender o desempenho de sistemas computacionais:***
  - *Escrever software de qualidade requer conhecimento do hardware.*
- ***Projetar usando soluções de hardware***
  - *Intel, AMD, IBM, ARM, Qualcomm, Apple, Oracle, NVIDIA*, Samsung, …

A Lenda das duas Arquiteturas

- **Computer architecture**
  - Definição da ISA (*Instruction Set Architecture* - Arquitetura de Conjunto de Instruções, em tradução livre) de forma a facilitar a criação de software;
  - Definição das interfaces entre software e hardware.
- **Computer micro-architecture**
  - Projeto do Processador, memória, estrutura de I/O;
  - Preocupação com a eficiência na interface entre software e hardware.

Projetos de Aplicação Específica

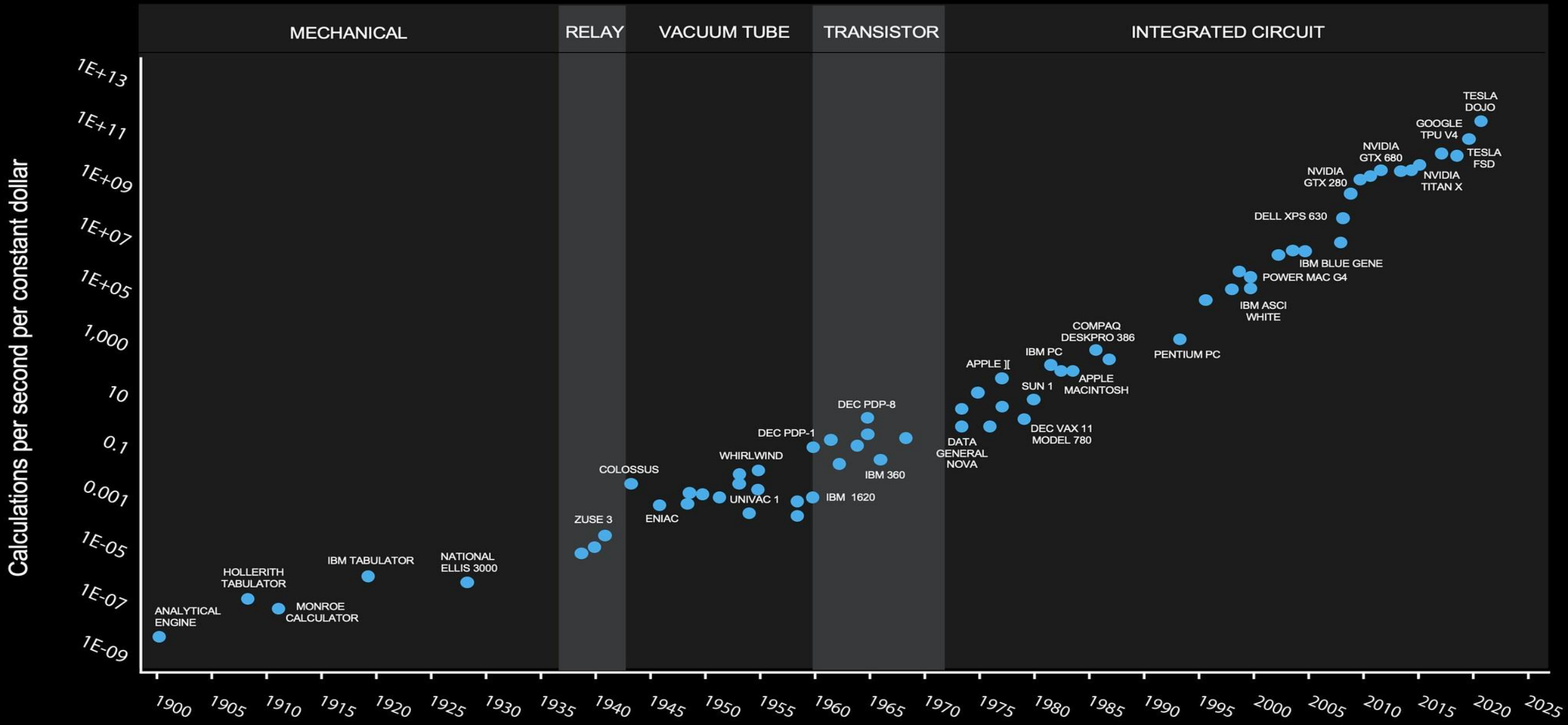
- **CPU's de Uso Geral**
  - O processador pode resolver qualquer problema;
  - Intel Atom/Core/Xeon, AMD Ryzen/EPYC, ARM M/A series.
- **Processadores de Uso Específico**
  - ASICs (Application specific integrated circuits – Circuitos integrados de aplicação específica)
    - Funcionalidades criticas de domínio específico;
    - Exemplos: video encoding, 3D, machine learning.
  - Em Geral
    - O hardware é pouco flexível;
    - Mais "paralelismo" vetorização, na verdade;

A Tecnologia Evolui

- Elementos básicos:
  - Transistores de estado sólido (chaveamento eletrônico);
  - Base da construção dos circuitos integrados.
- Caracteristicas dos circuitos integrados:
  - Alta desempenho e confiabilidade;
  - Baixo custo e consume;
  - Produção em massa.
- Famílias de circuitos integrados:
  - SRAM/logic: otimizada para velocidade e usada internamente nos processadores;
  - DRAM: otimizada para densidade, cuto e consume usada para armazenamento;
  - Flash: otimizada para densidade e custo, usada para armazenamento.

FUTURE VENTURES

122 YEARS OF MOORE'S LAW

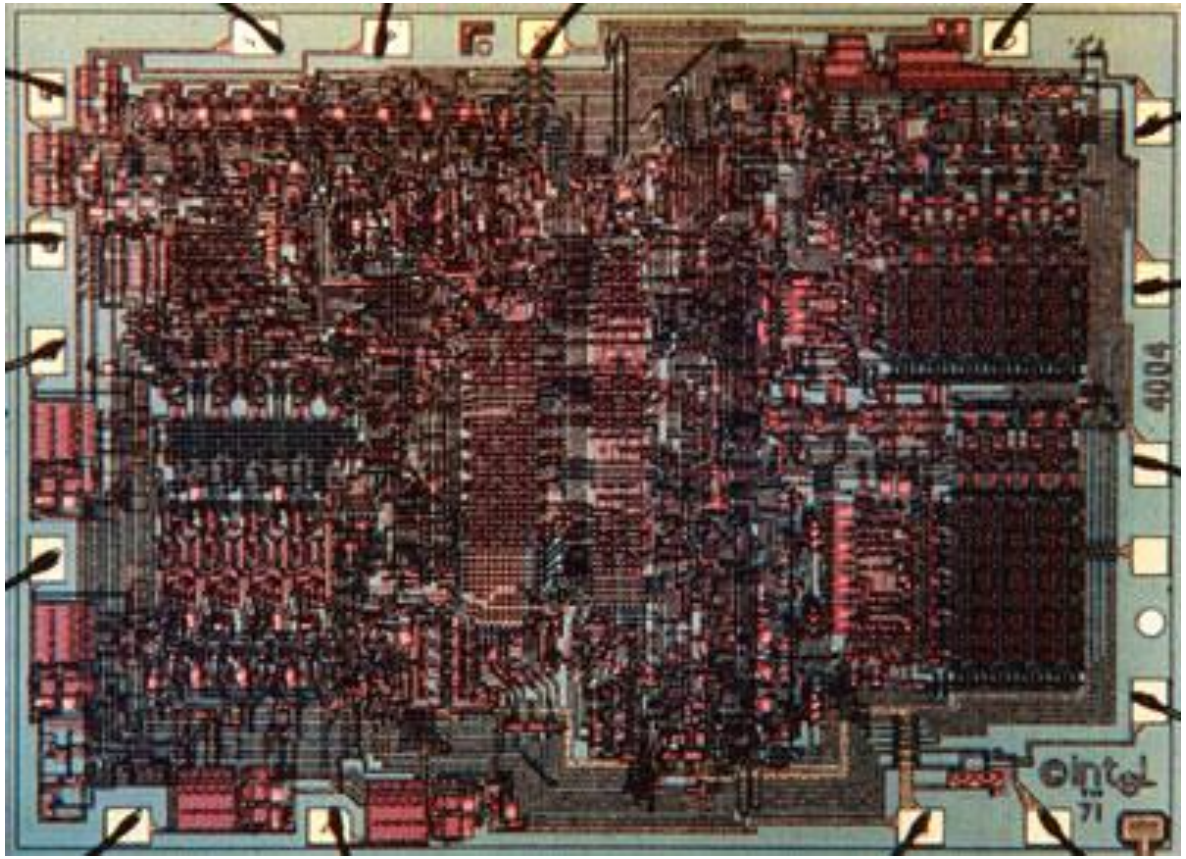


Tesla Has Done Something No Other Automaker Has: Assumed The Mantle Of Moore's Law – (evobsession.com)

Source: Ray Kurzweil, Steve Jurvetson

A Revolução do 4004

- Intel 4004 (1971)
  - Desenvolvido para calculadoras;
  - Tamanho do transistor: 10,000 nm
  - 2300 transistores
  - Área: 13 mm2
  - Clock: 108 KHz
  - Fonte:12 Volts
  - Barramento: 4-bit data



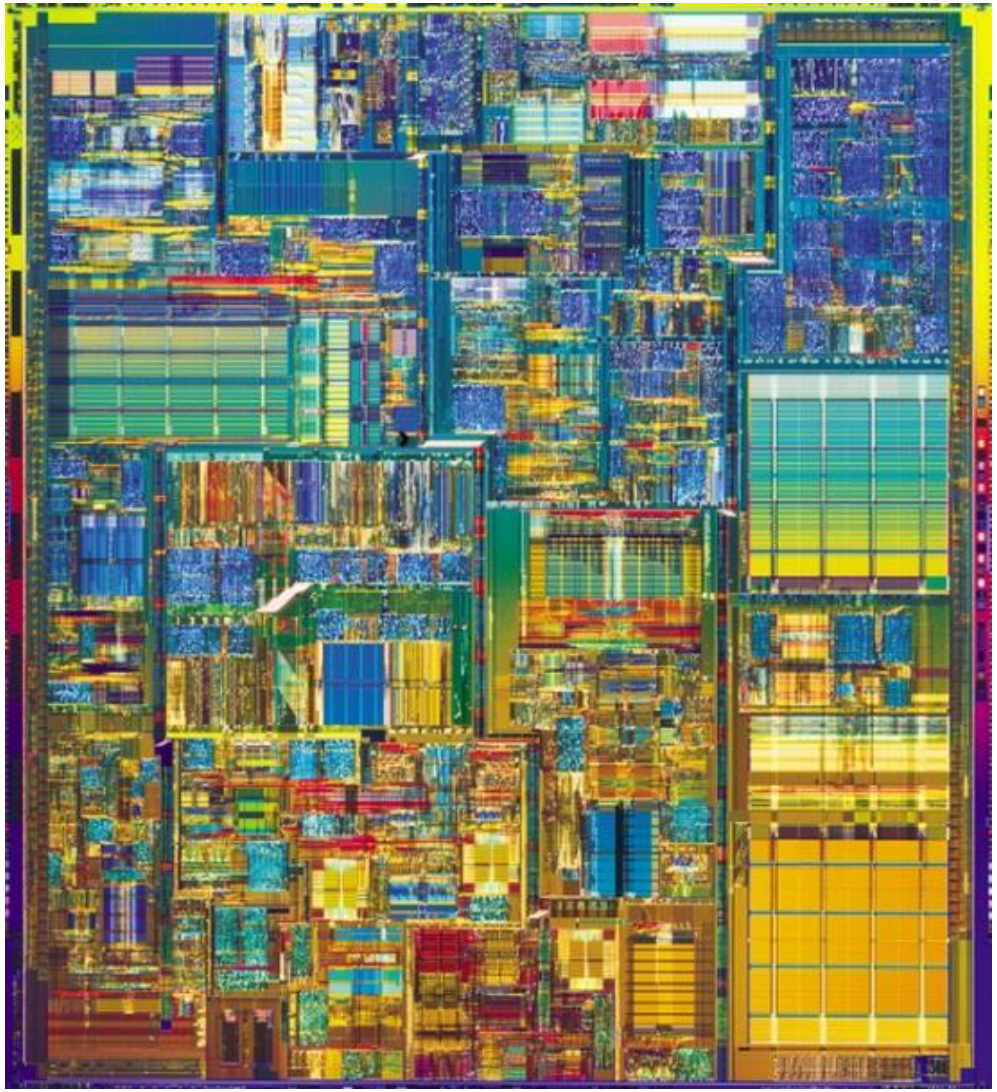
A Segunda Revolução: Paralelismo Inplícito

- Paralelismo implícito em nível de instrução:
  - O hardware fornece recursos que permitem o paralelismo de instruções;
  - Transparente para o software.

- Pipelining: a primeira técnica usada, além de fornecer operações em parelelo permitiu um aumento na frequência de clock.
- Caches: permitindo acesso a dados em paralelo. Forçado pelo aumento do clock;
- Processadores de ponto flutuante integrados;
- Mais pipelines, *branch speculation;* múltiplas instruções por clock; agendamento dinâmico.

Intel Pentium IV

- Intel Pentium4 (2003)
  - Aplicação: desktop/server;
  - Tamanho do Transistor: 90nm;
  - 55M transistores em área de 101 mm2;
  - Clock: 3.4 GHz;
  - Fonte: 1.2 Volts;
  - Barramento: 32/64-bit data (16x);
  - 22-stage pipelined datapath;
  - 3 instruções por ciclo;
  - Dois níveis de cache no processador;
  - Vetores de Dados (SIMD), hyperthreading

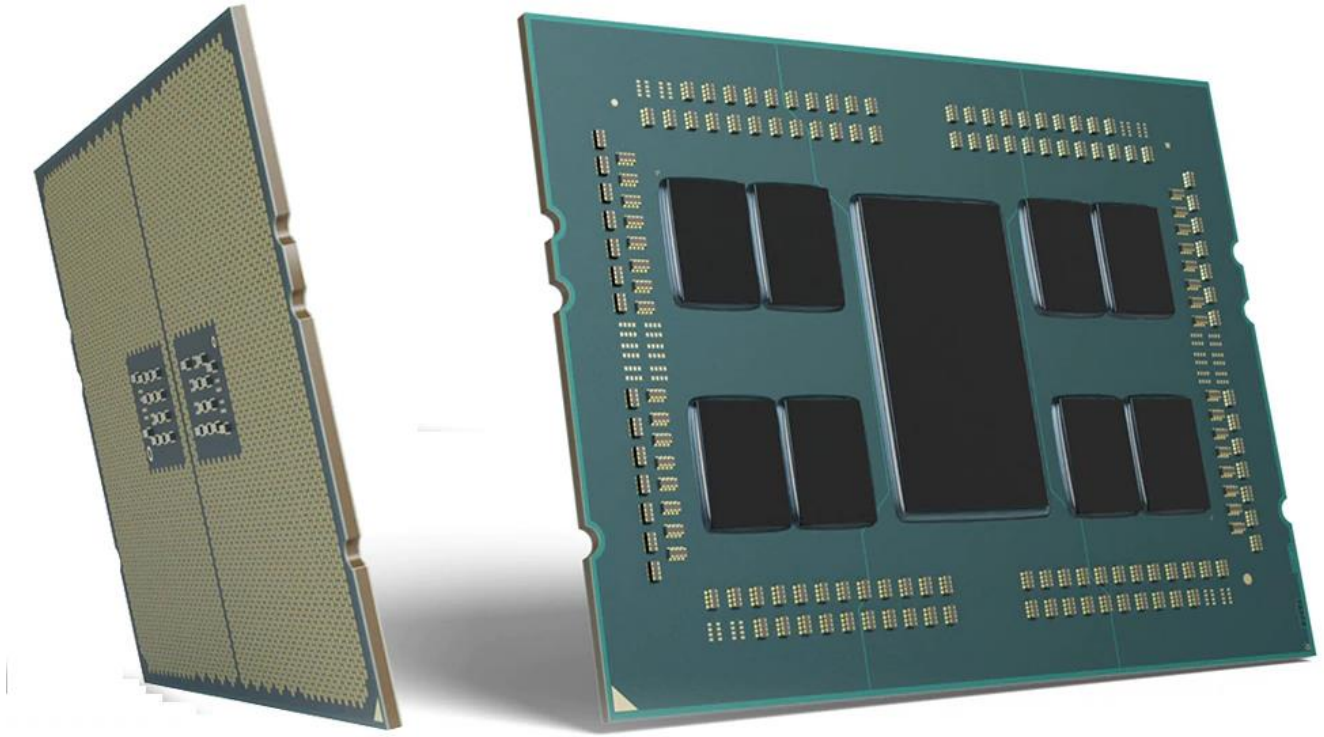


Terceira Revolução: paralelismo Explícito

- Explicitar o suporte a paralelismo de dados e *threads:*
  - O hardware disponibiliza recursos para uso em paralelo, o software determina o uso.
- Começamos usando instruções vetoriais, Intel's SSE, uma instrução faz 4 multiplicações em paralelo.
- Adicionamos suporte complete a multithread: caches coerentes, sincronização de hardware, etc.
- Depois adicionamos suporte para diversos *threads*, concorrentes, no mesmo dispositivo.
- Finalmente chegamos aos dispositivos contend múltiplos núcleos de processamento (multicore).
- Uma GPU é um processador com milhares de cores em uma arquitetura SIMD.

Processadores Multicore

- **AMD EPYC 7H12**
  - Aplicação: desktop/server;
  - Tamanho do transistor: 7nm;
  - 39.5B transistores;
  - Área: 1008 $mm^{2;}$
  - Clock: 2.6 to 3.3 Ghz;
  - Barramento: 256-bit data (2x);
  - 19-stage pipelined datapath;
  - 4 instruções por clock;
  - 292MB de cache on chip;
  - data-parallel vector (SIMD) instructions, hyperthreading;
  - **128 threads em 64-core multicore.**



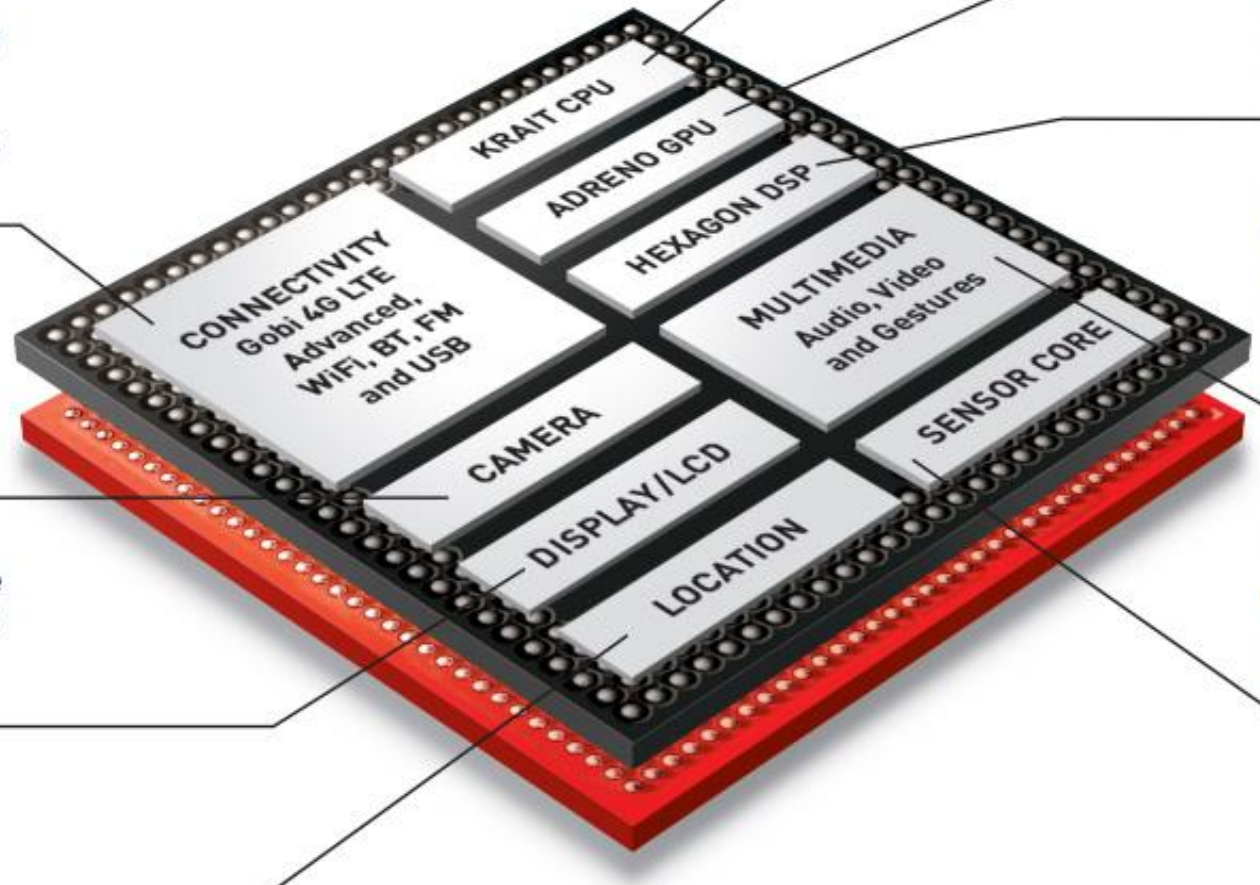
Para Comparar

	Intel 4004	Intel Pentium 4 Prescott	AMD EPYC 7H12
Lançamento	1971	2004	2019
Tamanho do Transistor	10,000 nm	90 nm	7 nm, 14 nm
Número de Transistores	2,300	125M	39.5B
Área	13 mm^2	112 mm^2	1008 mm^2
Clock	740 KHz	3.8 GHz	2.6-3.3 GHz
Barrramento	4-bit	64-bit	256-bit
Estágios de Pipelining	n/a	31	19
Largura da Pipeline	n/a	3	4
Número de Núcleos	1	1	64
Cache	n/a	1MB	292MB

Quarta Revolução: aceleradores

- Combinação de tipos diferentes de núcleos de processamento no mesmo chip:
  - System-on-Chip (SoC) em sistemas embarcados e dispositivos móveis.
  - Graphics Processing Units (GPUs)
  - media codecs, radios, encryption, compressão, machine learning
- Desempenho excelente, consume excelente: programação extremamente complicada.

SNAPDRAGON 805 PROCESSOR



Faster performance and more multitasking with Krait 450 CPU at up to 2.7 GHz

Console quality gaming with new generation Adreno 420 GPU

More power-efficient apps and system processing with the Hexagon™ QDSP6

Capture and play back Ultra HD video and enjoy 7.1 surround sound on the go or at home with advanced video and audio engines

Get more use and greater accuracy from sensor-intensive apps with the dedicated Snapdragon Sensor Engine

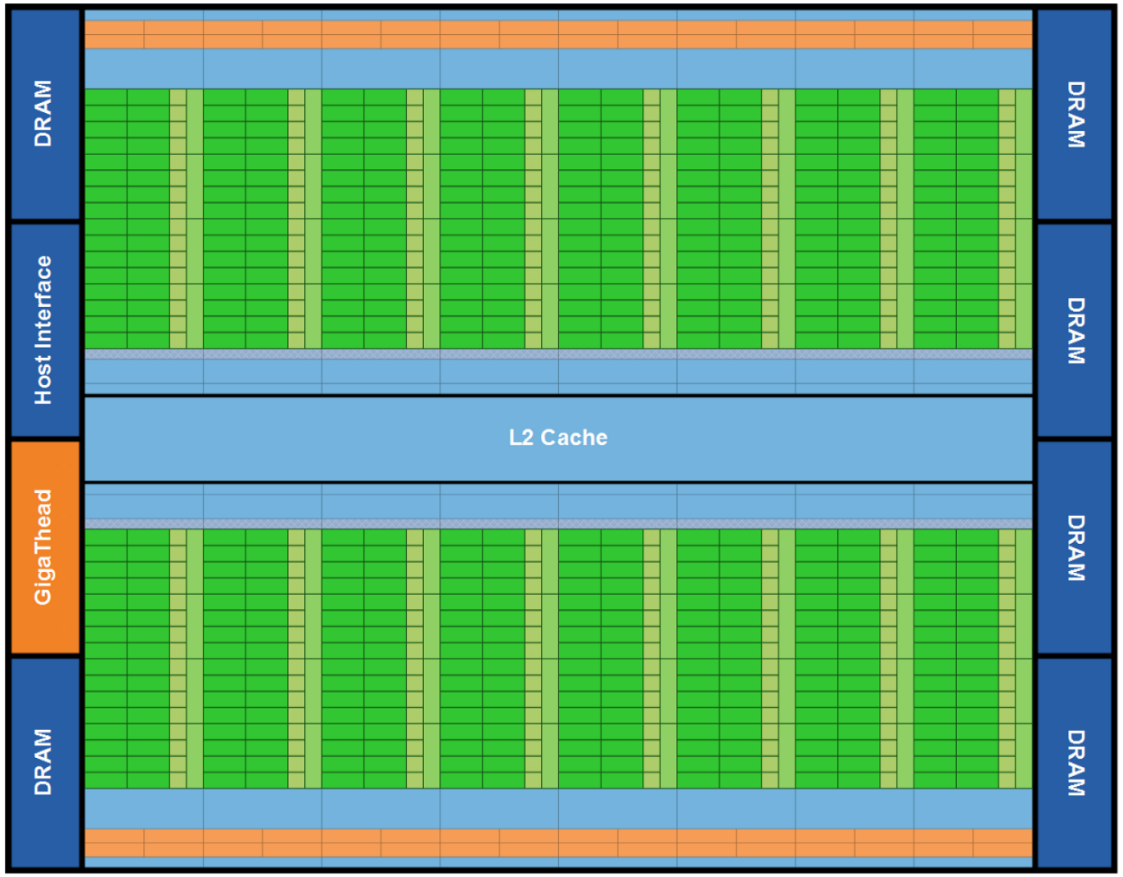
Stay connected and stream large files fast with industry leading connectivity, including the world's most advanced 4G LTE and VIVE™ 2-stream 802.11ac Wi-Fi

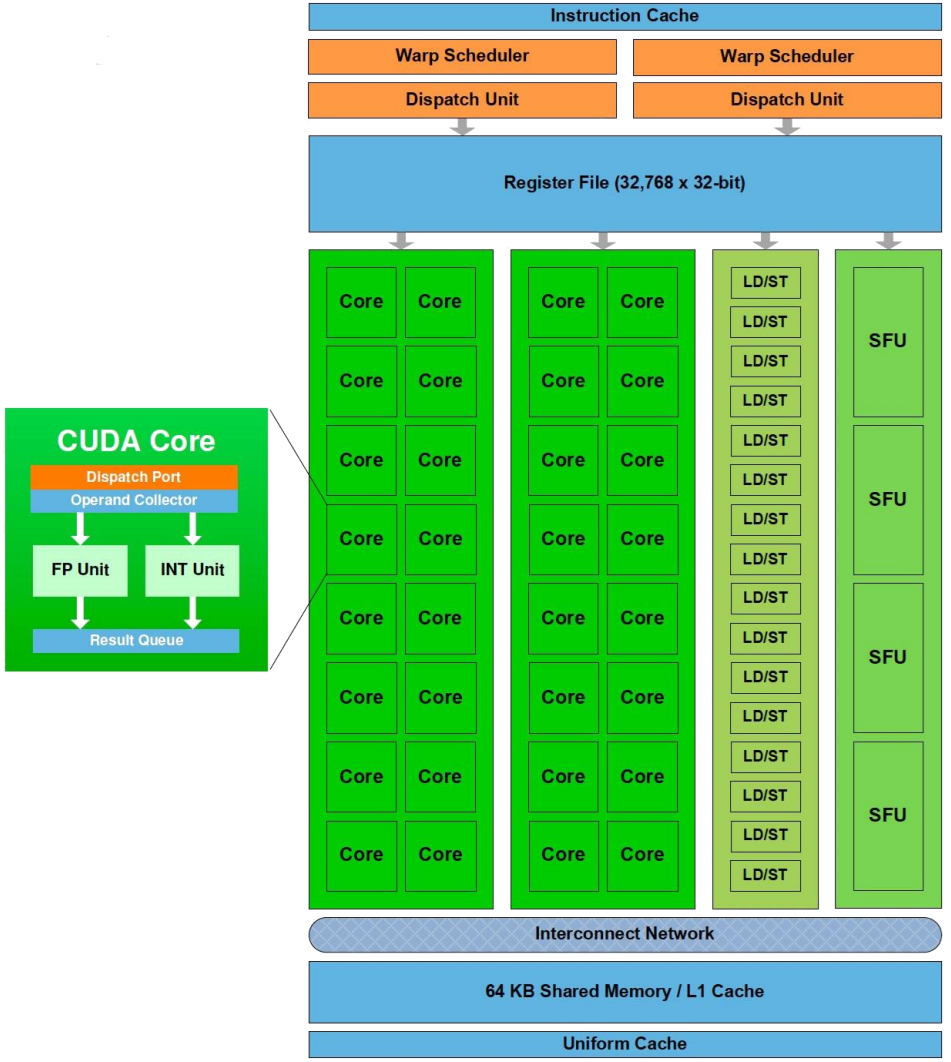
Capture sharper photos, even in low light, with the mobile industry's first dual ISP

Enjoy Ultra HD resolution content on Ultra HD-capable mobile devices and Ultra HD TVs with the Snapdragon Display Processor

Find your way outdoors and indoors with IZat GNSS with support for GPS, Glonass and BeiDou constellations

Nvidia - Fermi





Fermi Streaming Multiprocessor (SM)

Fonte: Nvidia (2020)

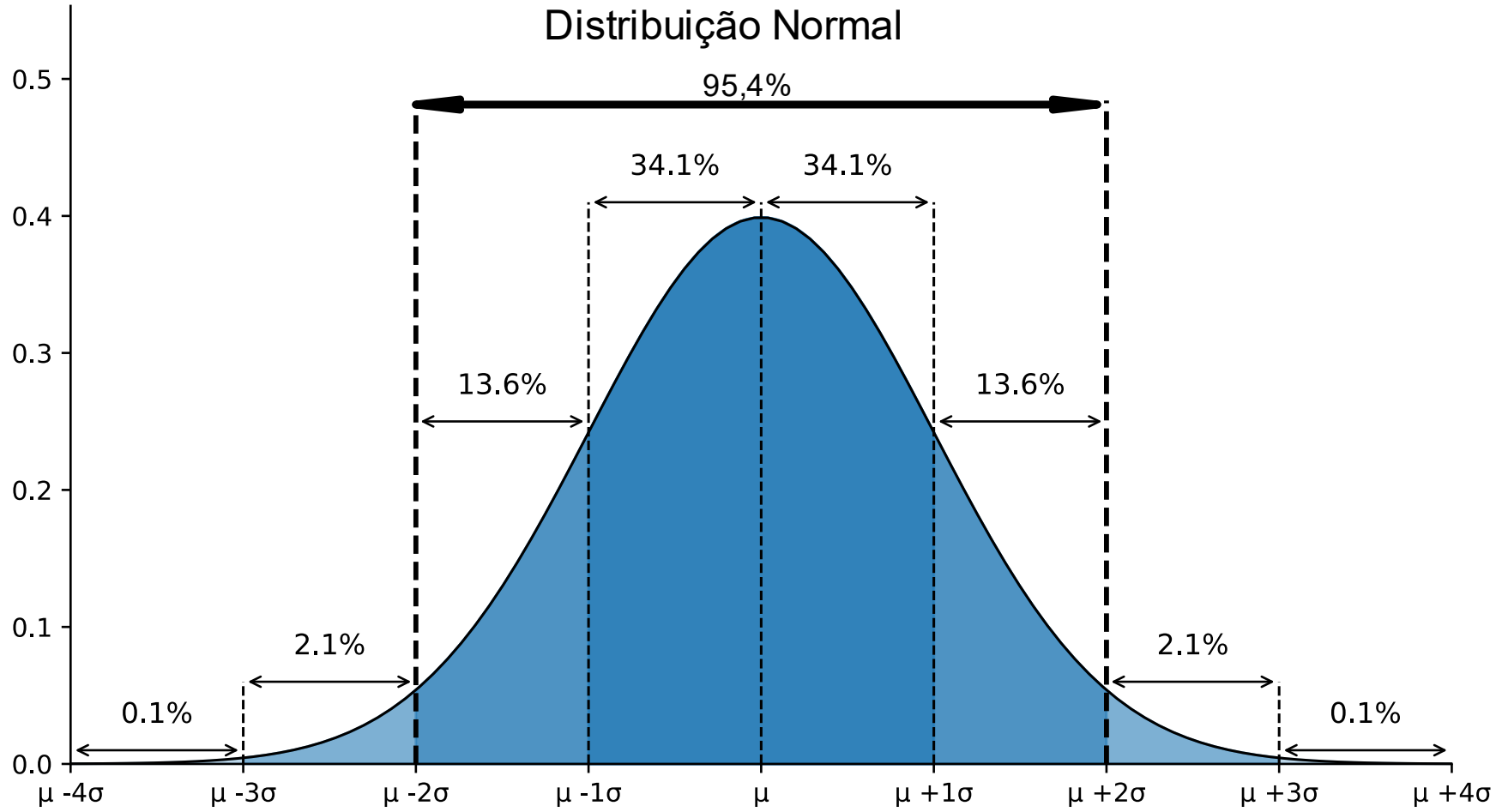
Desempenho

Computer science is no more about computers
than astronomy is about telescopes.
Edsger Dijkstra

Conceito

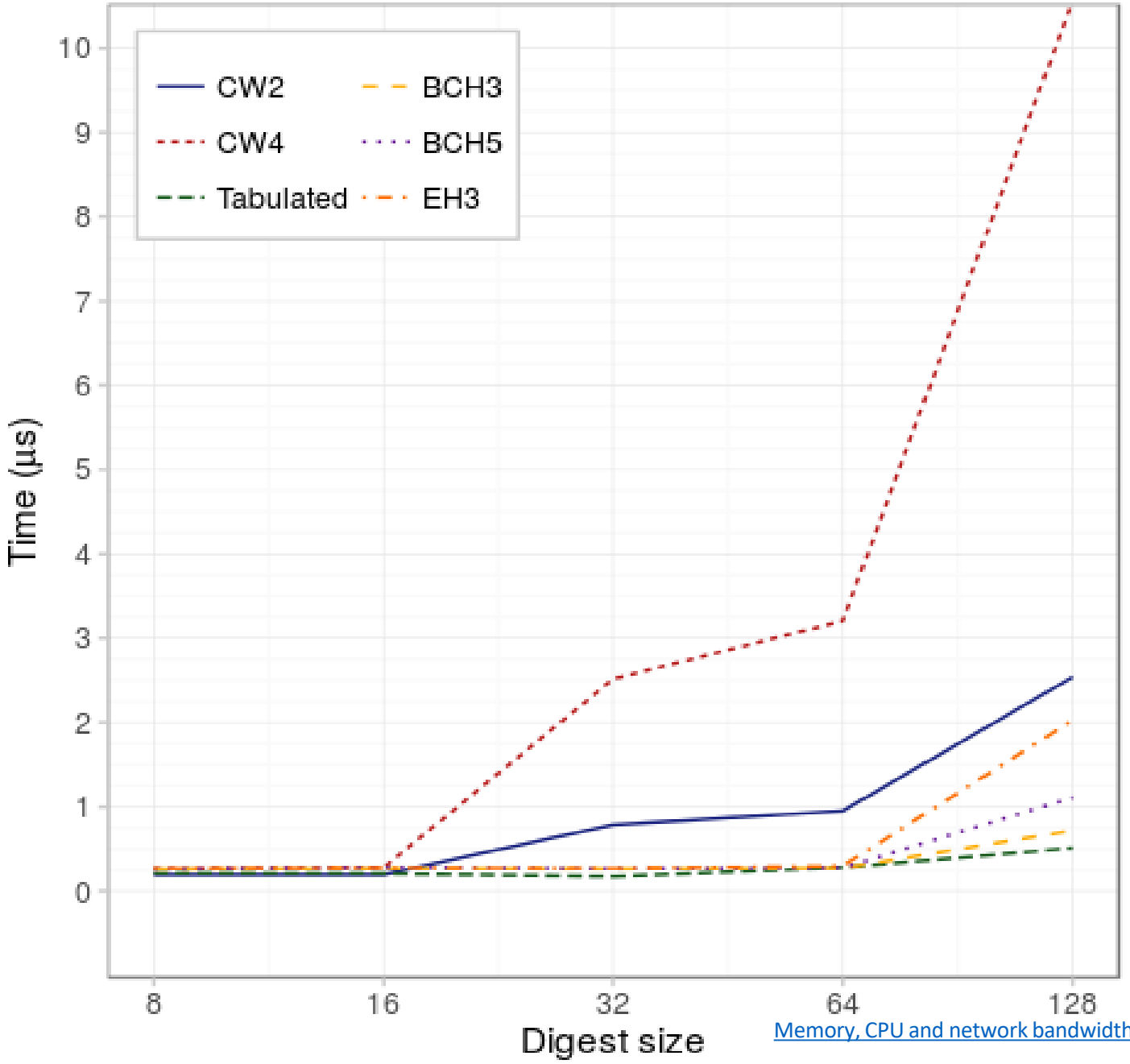
- Desempenho, ou performance, é uma medida da eficiência e rapidez com que um sistema computacional ou programa executa uma determinada tarefa.
- É uma medida de quão bem um sistema ou programa está operando em relação às suas especificações;
- É uma medida ambígua de qualidade que, geralmente, está relacionada com a velocidade, eficiência e utilização de recursos, como CPU, memória, armazenamento e rede.

Distribuição Normal



Fonte: o autor (2022)

Geração de Números Pseudoaleatórios

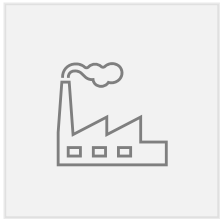


Memory, CPU and network bandwidth costs (esterl.github.io)

Quatro Custos

Custo de Desenvolvimento: o custo do processo de design, criação e codificação do produto.

Custo de Tradução: o custo de traduzir a linguagem de programação para linguagem de máquina.



Custo de Produção: o custo para manter a solução desenvolvida rodando e atendendo clientes.

Custo de Manutenção: o custo de correção de erros e de evolução do sistema.

Atividade Prática

A prática é o melhor professor.
Marcus Antonius

Atividade Prática – Em Grupo

- Usando um ambiente *online* (repl.it ou Google Colaboratory), escreva um código em Python que permita medir o espaço de memória e o tempo necessário, em nano segundos, às seguintes funções: (a) Criar um array com 1000 caracteres aleatórios; (b) Ordenar um array com 1000 caracteres; (c) Imprimir, no terminal um array com 1000 caracteres; (d) calcular π Com 100 casas decimais usando a aproximação de Leibnitz:

$$\pi = 4 \cdot \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1}$$

Obrigada!

Frank de Alcantara
frank.alcantara@pucpr.br

SCIENTIA, VITA ET FIDES
PUCPR
GRUPO MARISTA