

EGM0017 (60h)

# Fluxo e metodologias de projeto de Sistemas Embarcados

**Prof. Josenalde Barbosa de Oliveira – UFRN**

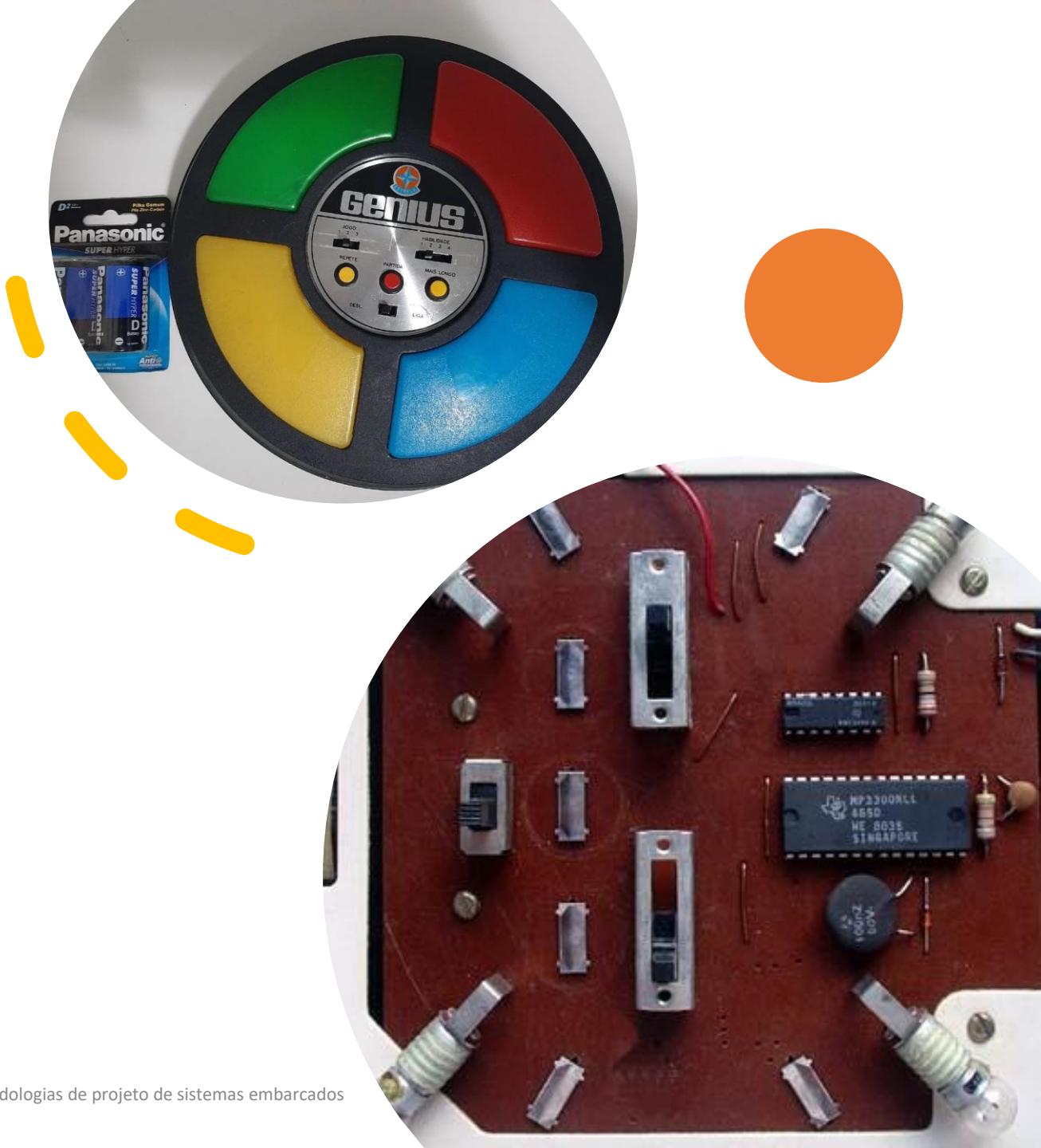


josenalde.oliveira@ufrn.br

Programa de Pós-Graduação em Engenharia Mecatrônica

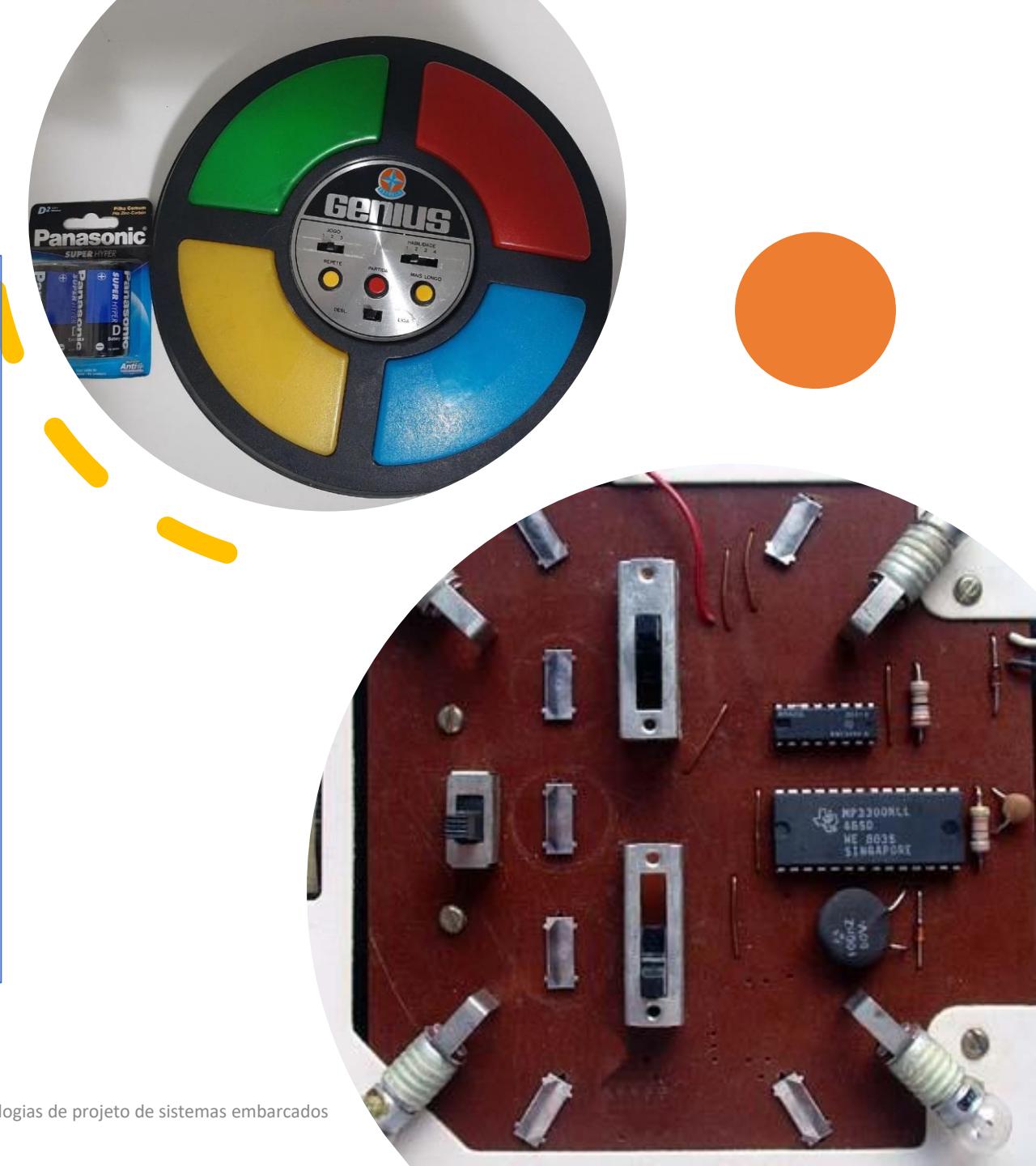
# Intro aos sistemas embarcados

- Sistema Embarcado (firmware+hardware)
  - construído especificamente para uma aplicação
  - Ex: Brinquedo eletrônico
  - uC: MP3300 (Texas) com Driver SN75494
  - Patenteável (INPI)
  - Requisitos FUNCIONAIS:
    - RF1. Emitir tons diferentes ao piscar cada led/lâmpada ou quando o usuário pressiona tecla
    - RF2. Em cada iteração, repetir a sequência de tons/cores anteriores e acrescentar 1 tom/cor aleatoriamente
    - RF3. Permitir repetir a sequência mais longa através de botão ao fim do jogo
    - RF4. Permitir repetir a última sequência jogada através de botão ao fim do jogo
    - RF5. Permitir escolher entre 3 tipos de jogo e 4 níveis de habilidade
    - RF6. Sinalizar fim de jogo com tom grave e 4 leds piscando 3x com duração 1s cada



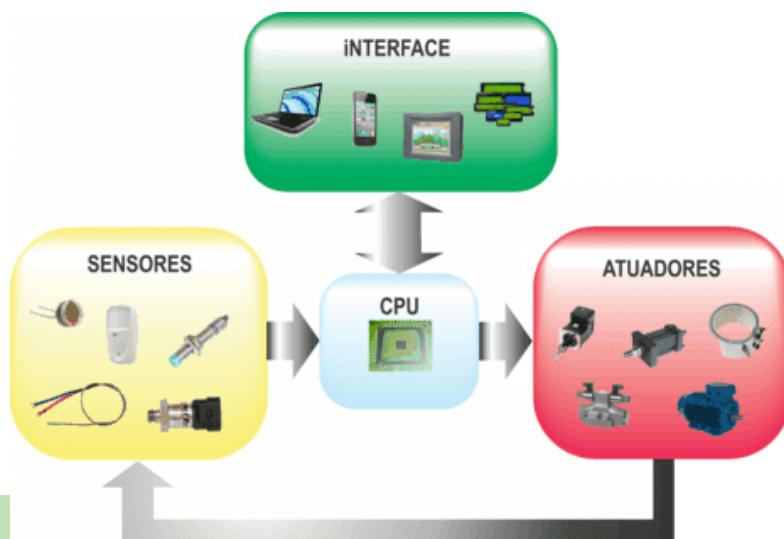
# Intro aos sistemas embarcados

- Requisitos NÃO FUNCIONAIS:
  - RNF1: Tempo de sequência em cada nível de habilidade (1:  $\leq 10s$ ; 2:  $\leq 8s$ ; 3:  $\leq 5s$ ; 4:  $\leq 3s$ )
  - RNF2: Tempo espera por toque  $\leq 5s$
  - RNF3: Alimentação 6V com consumo máximo de corrente: ?
- Regras de Negócio:
  - RN1: Tecla não pressionada em RNF2 ou erro na sequência – RF6
  - RN2: No Jogo tipo 2, emitir apenas 1 sinal e o usuário poderá criar sequência de até 31 sons no nível H4. Se RNF2 ou erro, RF6



# Intro aos sistemas embarcados - definições

- Combinação de hardware e software, com a possibilidade de integração de partes mecânicas e outras, projetada para realizar uma função dedicada; Podem fazer parte de um sistema ou produto maior, como exemplo o sistema de freios ABS de um carro
- Requisitos e restrições variáveis (segurança, confiabilidade, tempo real, flexibilidade, suporte à radiação, vibrações, umidade, tempo de resposta, precisão, etc).



## CATEGORIAS

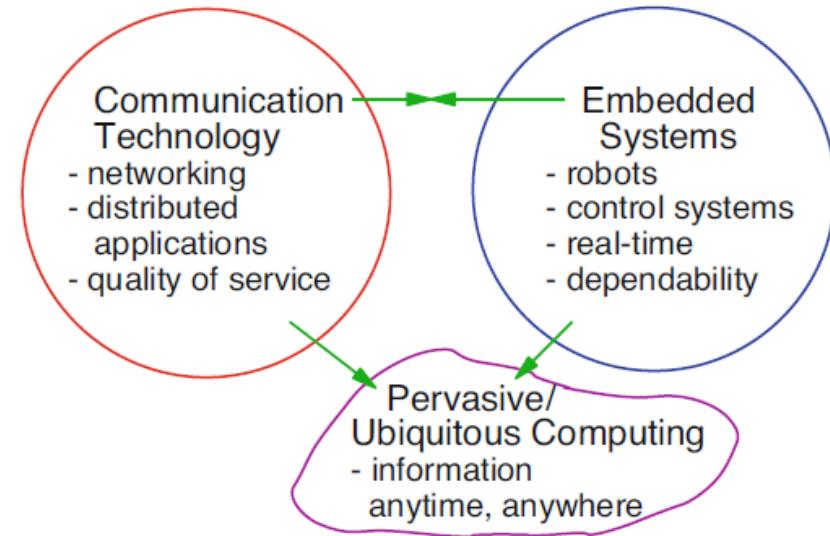
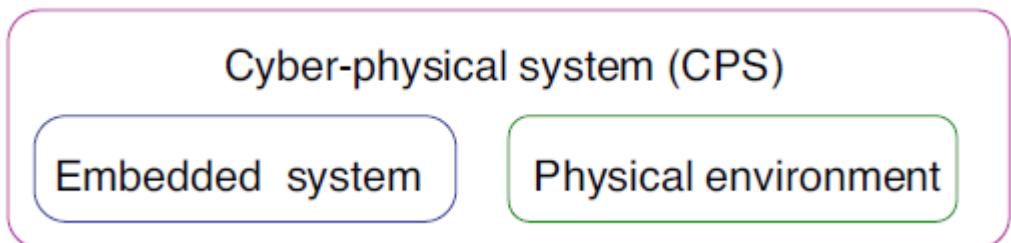
- 1) **TEMPO REAL/CRÍTICOS:** aplicações de armazenamento, automotivas, industriais, transporte, redes etc. (Free RTOS etc.)
- 2) **PLATAFORMAS:** dispositivos com SO abertos, como Linux, Android, Chrome etc.
- 3) **SEGURANÇA:** smart cards, placas SIM, terminais de pagamento



Quais os requisitos não funcionais de resistência/durabilidade neste ambiente?  
Exemplo: [apicultura e embarcados](#)

# Intro aos sistemas embarcados - definições

O conceito de **cyber sistemas** é uma extensão quando se tem necessariamente a interação com dispositivos físicos e **comunicação**



## Tecnologias habilitadoras 4.0

Cidades, Saúde, Indústria, Agro, Educação etc.

- Robótica, IA
- IoT
- Manufatura Aditiva (Fasten)
- Simulação
- Ciber-Segurança
- Big Data, Analytics
- Realidade Aumentada (Massive)

Laboratório de Computação Móvel e Ubíqua LabCOMU@IMD.

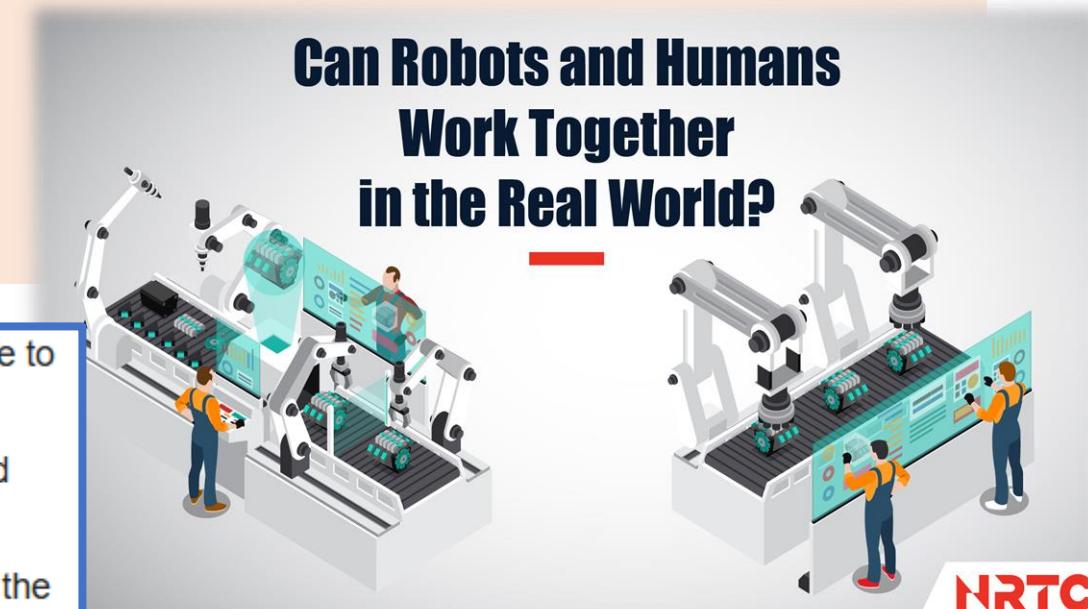
# Intro aos sistemas embarcados - definições

Alguns desafios neste tipo diferenciado de desenvolvimento de sistemas:

1. A confiabilidade exigida pela integração de cyber sistemas a IoT origina-se pelo fato dos sistemas estarem integrados a atuadores e dispositivos finais físicos, não causando dano(s) a operadores e outros sistemas

1. **Segurança externa:** preservar confiança, integridade e disponibilidade da informação. Logo, *IoT pode abrir portas?*
  1. *Estratégias de autenticação (usuário, dispositivo!)*
  2. *Camadas de proteção aos dados*
  3. *Níveis de acesso (restrições) – controle rigoroso*
  4. *Confidencialidade (criptografia)*
2. **Segurança funcional (interna):** *sistema/equipamento operar corretamente e no tempo certo em resposta às entradas; falhas de hw, sw, energia, operadores (humano)*

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



# Intro aos sistemas embarcados - definições



<https://owasp.org/www-project-embedded-application-security/>

**Alguns desafios neste tipo diferenciado de desenvolvimento de sistemas:**

**3. Disponibilidade:** probabilidade de um sistema estar disponível. Ela pode ser reduzida pelo defeito em componentes *em operação (falha)*. Está diretamente relacionada à ‘reparabilidade’, ou seja, a probabilidade um sistema ser recuperado em certo tempo. Dada por MTBF / (MTBF+MTTR)

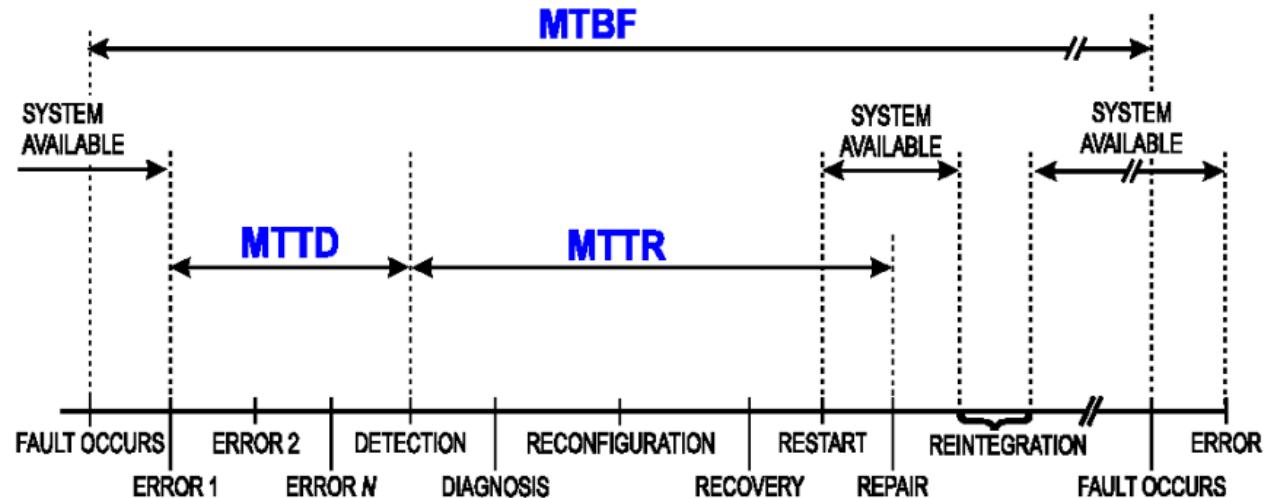
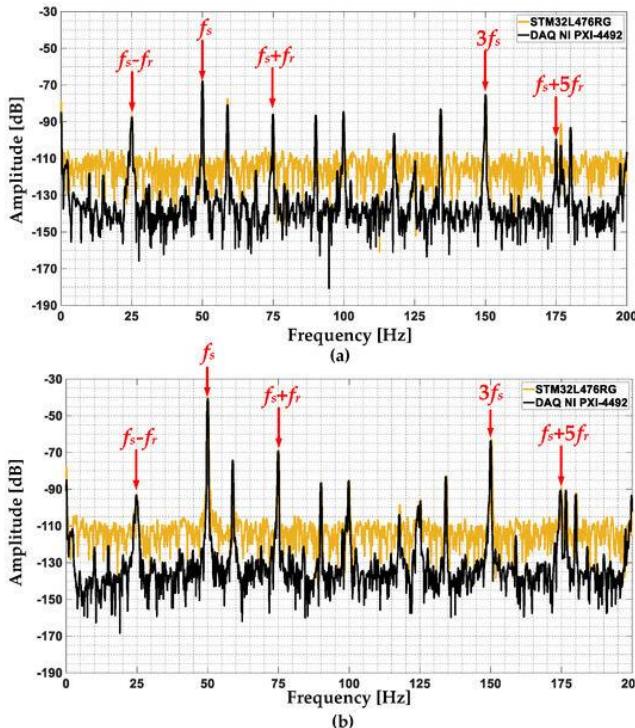
**MTBF (Mean Time Between Failures)** refere-se à quantidade média de tempo que um dispositivo ou produto funciona antes de falhar

**MTTR (Mean Time To Repair)** medida da capacidade de manutenção de um item reparável, que informa o tempo médio necessário para reparar um item ou componente específico e retorná-lo ao status normal de trabalho (notificação, diagnóstico, reparo real).

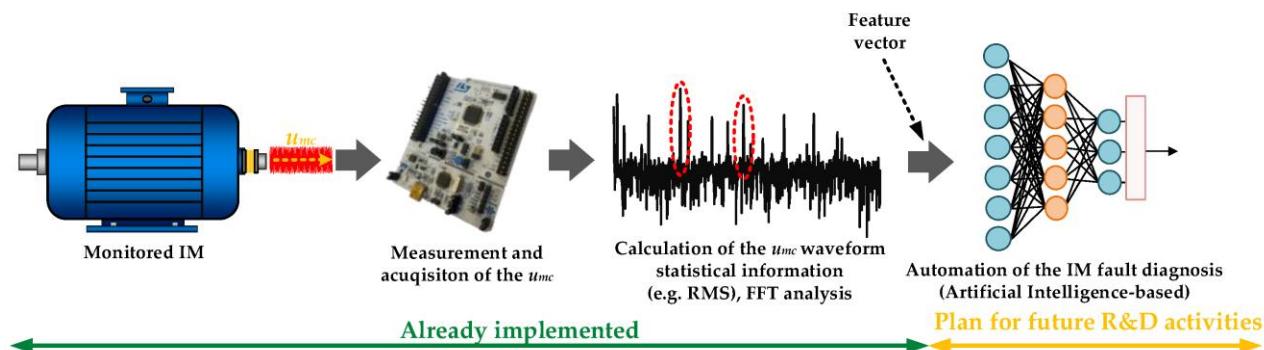
# Intro aos sistemas embarcados - definições



Exemplo prático!



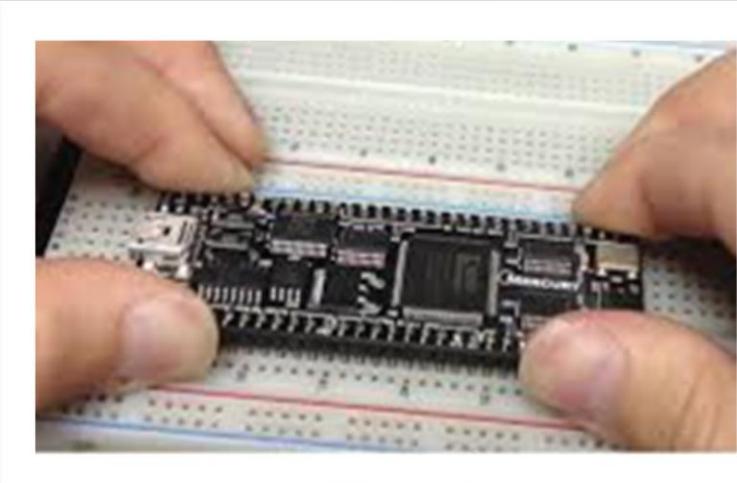
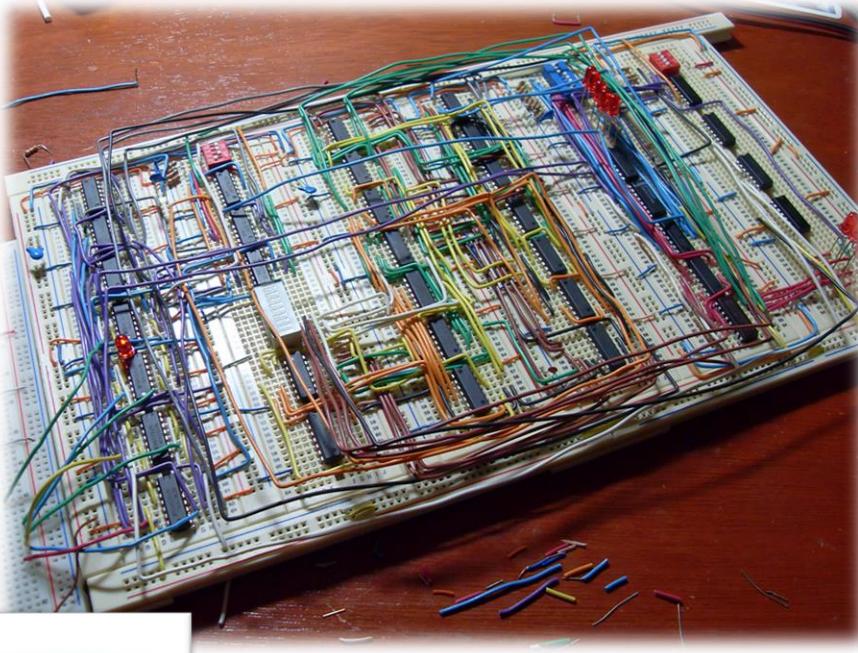
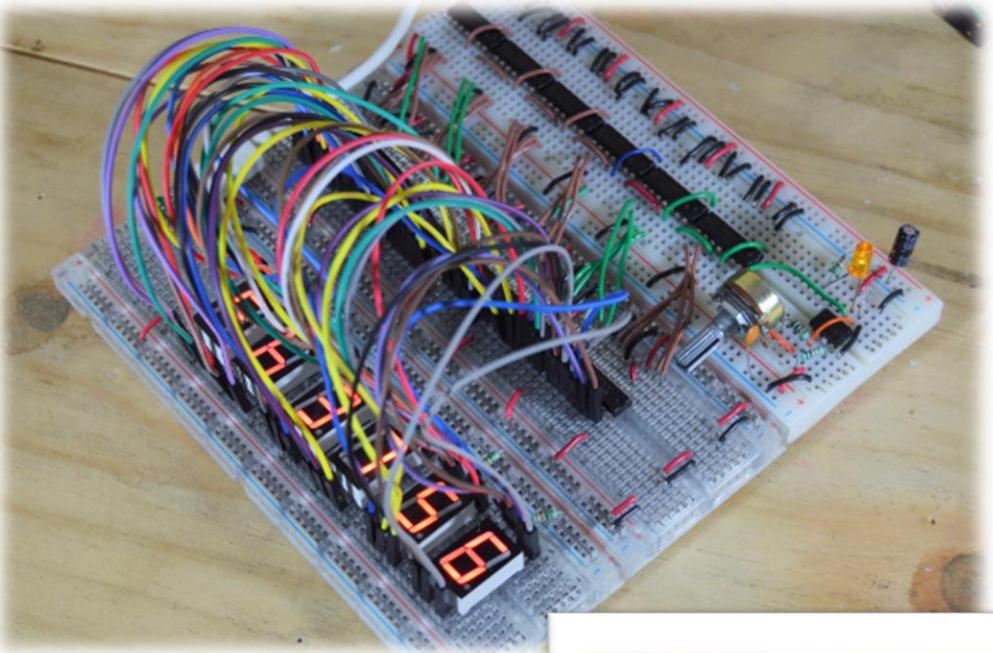
A Scenario for on-line detection and off-line repair. The measures – MTBF, MTTD, and MTTR are the average times to failure, to detection, and to repair.



- [1] Pietrzak et al. (2024)  
[2] Oliveira et al. (2017)

Tolerância, Detecção, isolamento,  
Recuperação de falhas: área em  
crescimento para sistemas embarcados com IA

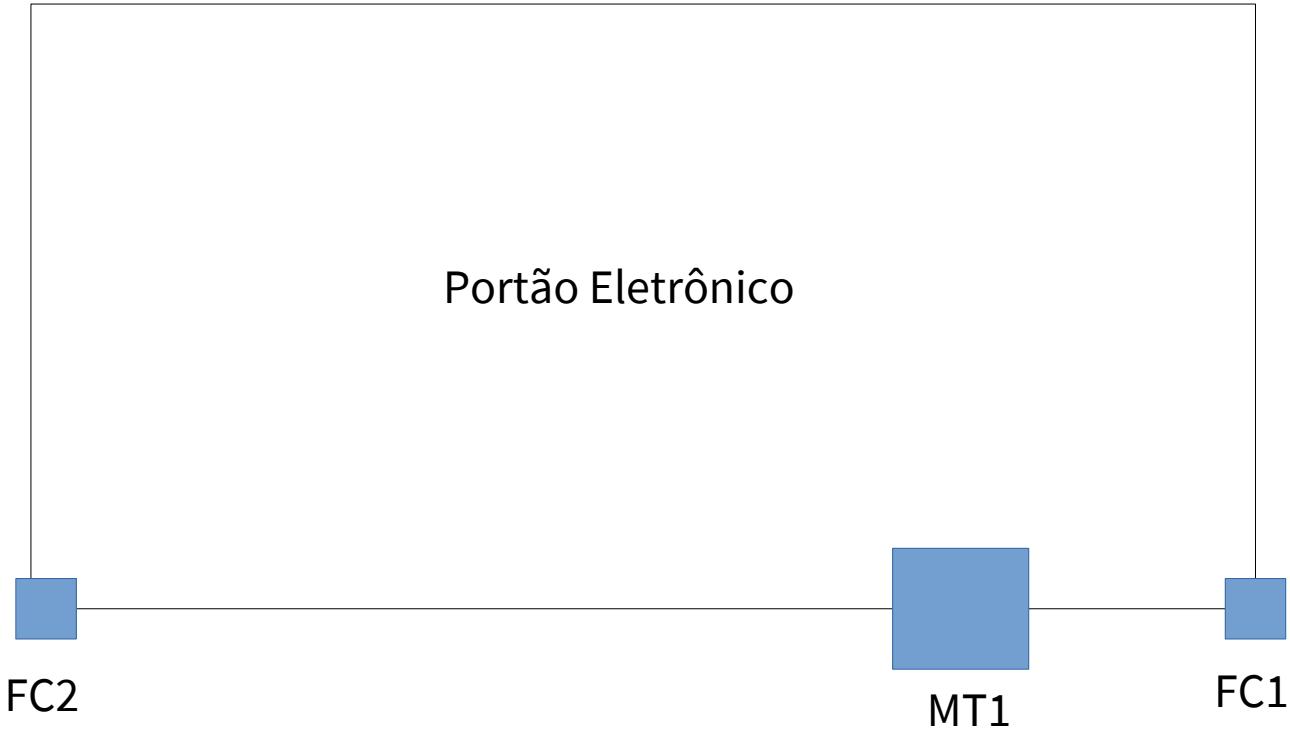
Mas...existem variedades de soluções  
Como escolher?



Que tal esta solução?

# Existem variedades de soluções

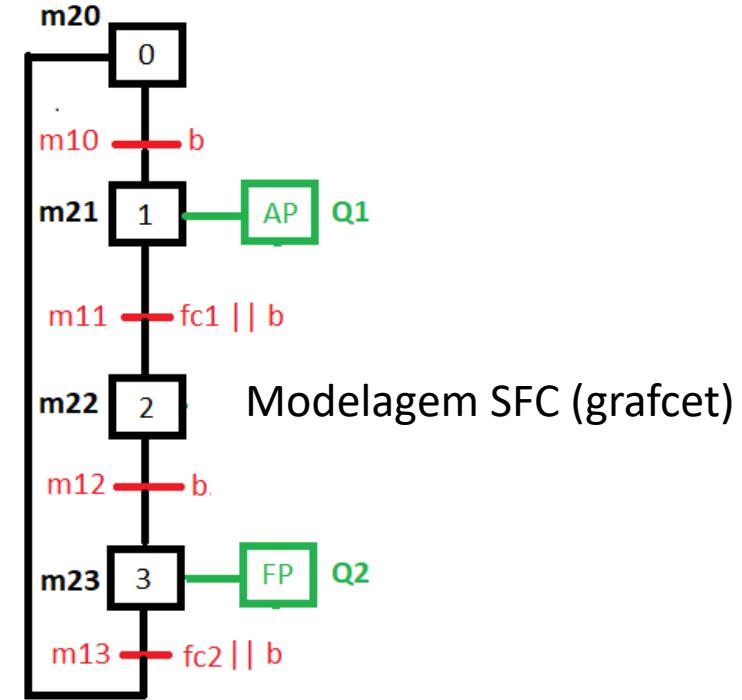
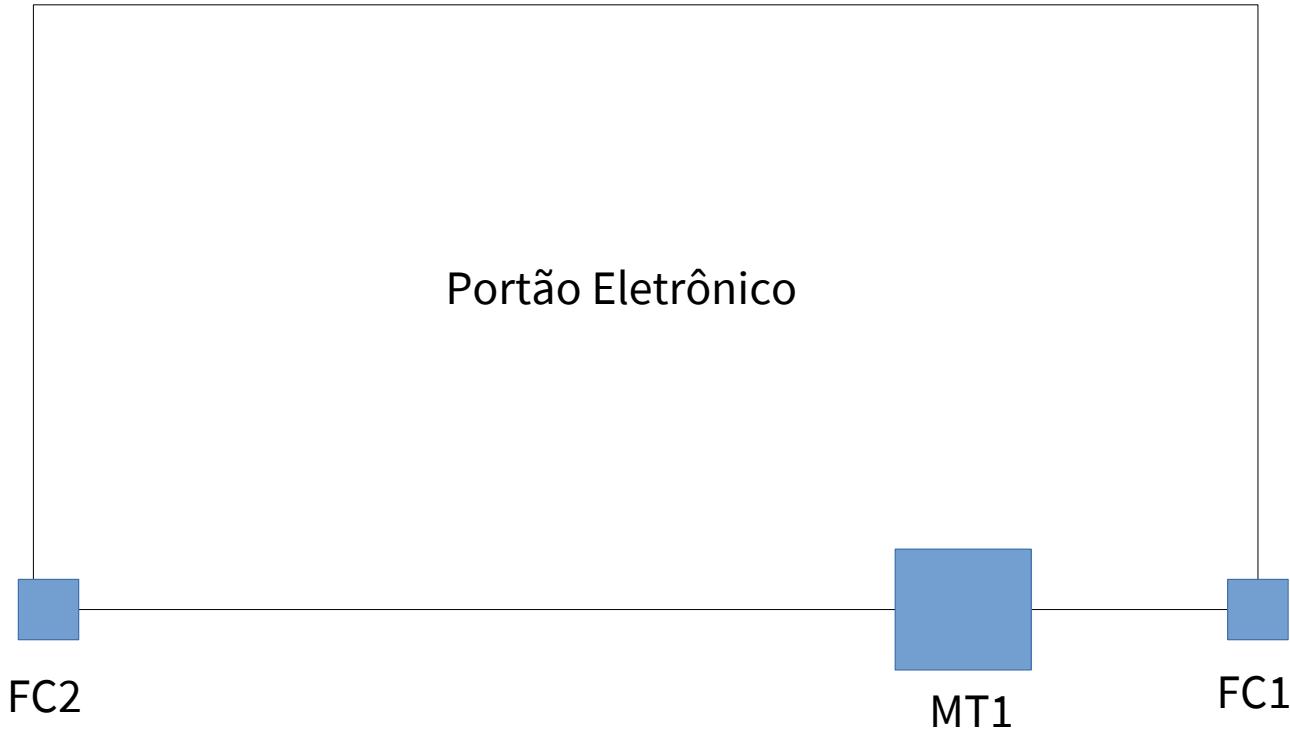
## Exemplo : Entradas e Saídas Digitais



Como programar uma solução aqui?

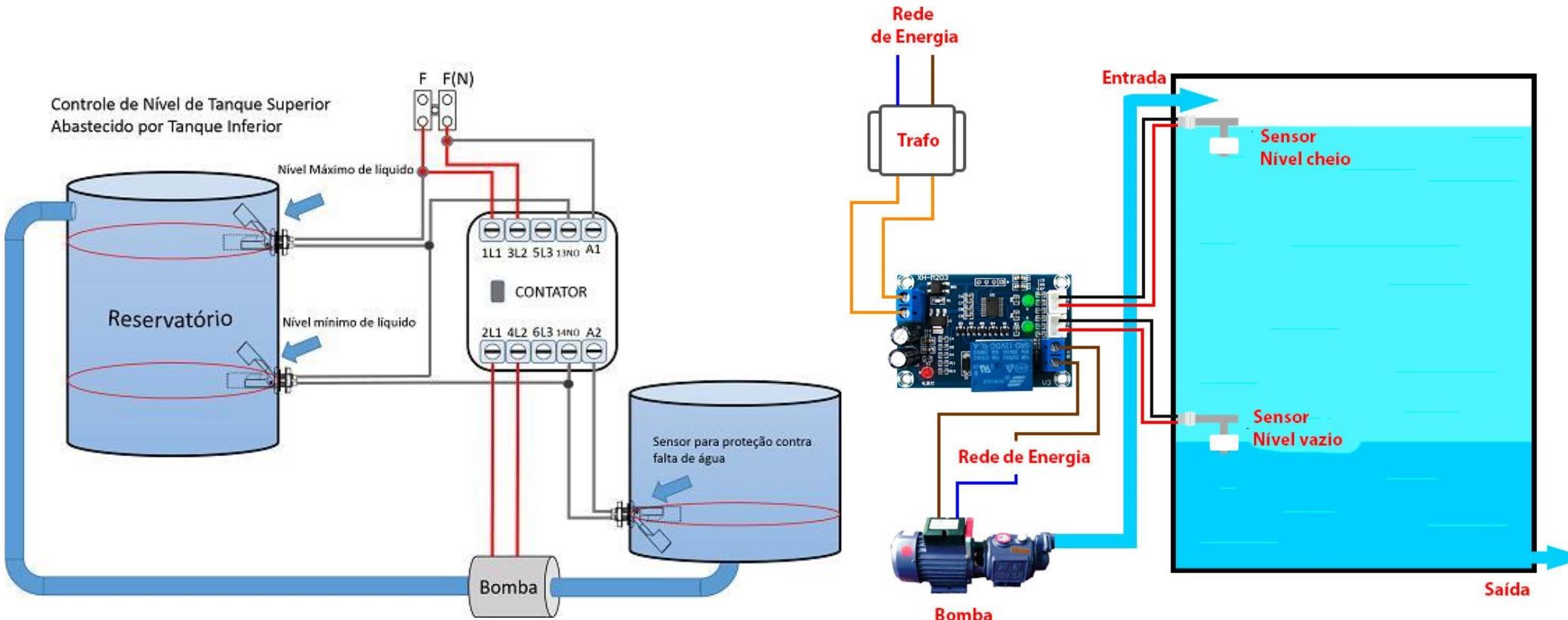
# Existem variedades de soluções

## Exemplo : Entradas e Saídas Digitais



# Existem variedades de soluções

## Exemplo : Entradas e Saídas Digitais: Controle de Nível



# Existem variedades de soluções



**Sistema de Controle para Transportador Automático de Peças**  
**Objetivo: transportar peças/objetos e retirá-los automaticamente**

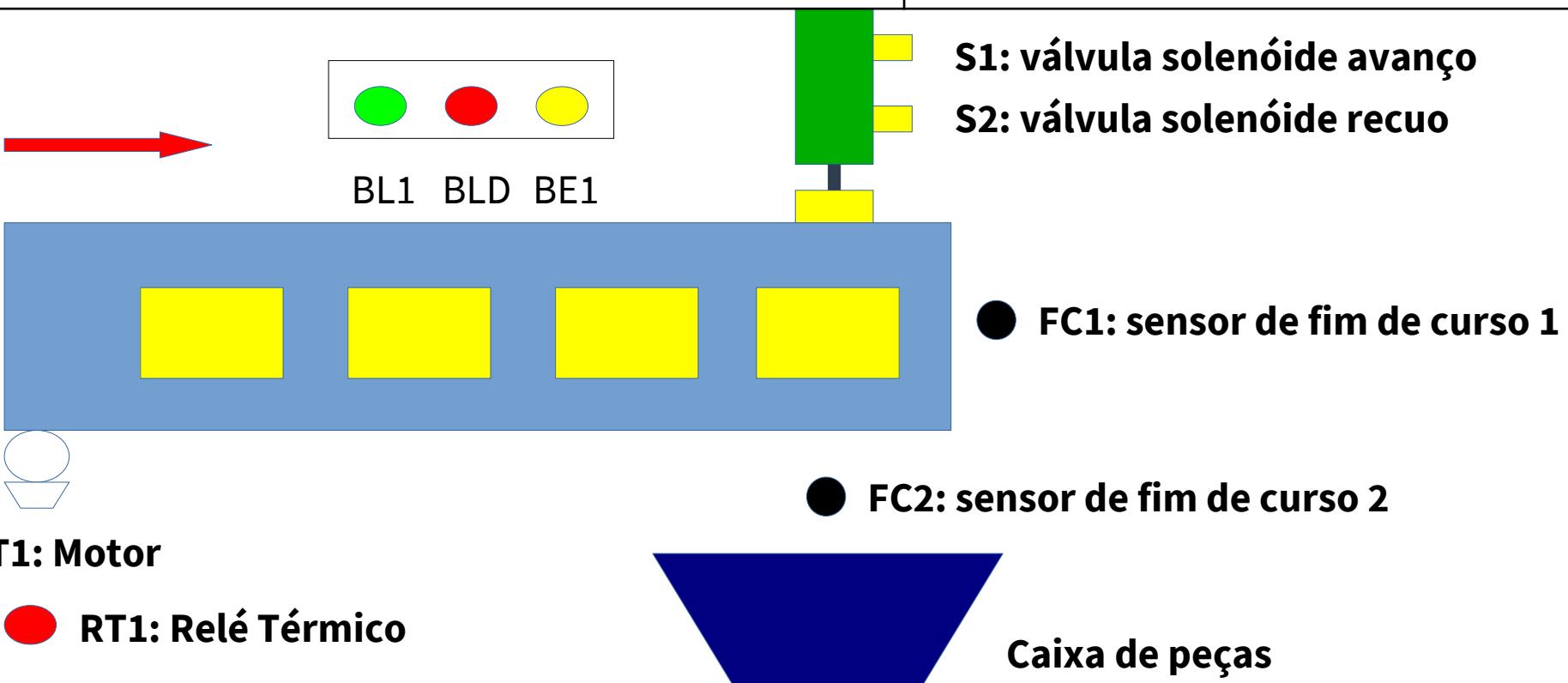
# Existem variedades de soluções

## Entradas:

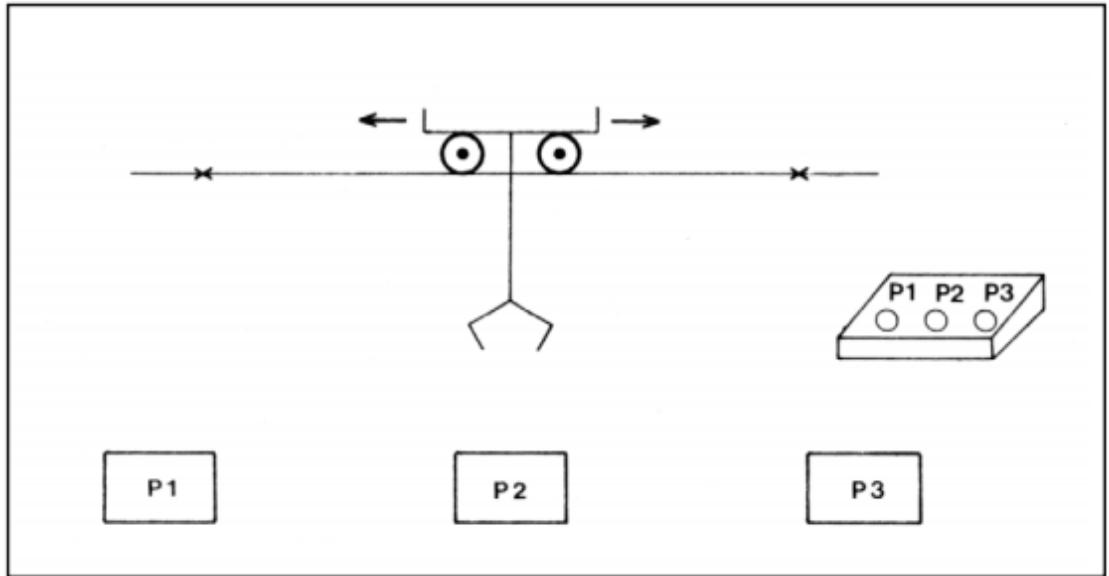
BL1: botoeira para ligar o motor  
BLD: botoeira para desligar o motor  
BE1: botoeira de emergência  
RT1: relé térmico (protege o motor em caso de curto circuito)  
FC1: sensor de fim de curso 1 (peça pronta para ser retirada da esteira)  
FC2: sensor detecção peça fora da esteira

## Saídas:

MT1: motor  
S1: solenóide avança pistão  
S2: solenóide recua pistão  
L1: sinaliza motor desligado (BLD)  
L2: sensor FC1 acionado  
L3: sensor FC2 acionado  
L4: sinaliza relé térmico acionado  
L5: sinaliza BE1 acionado



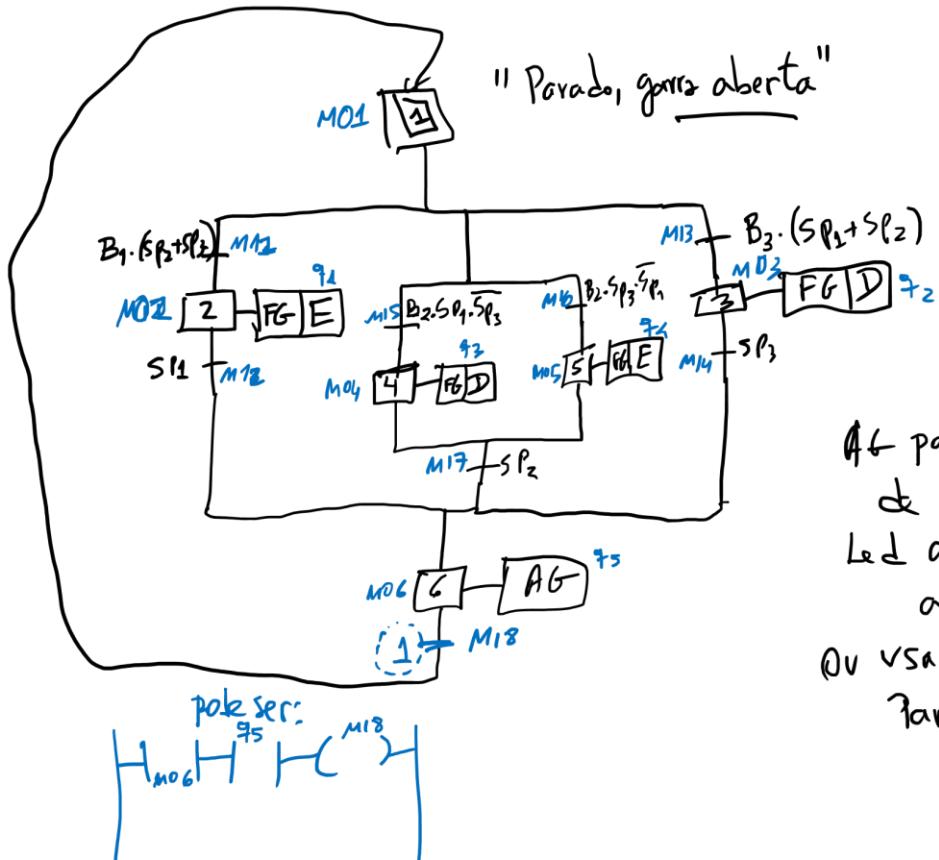
# Problema robô-garra



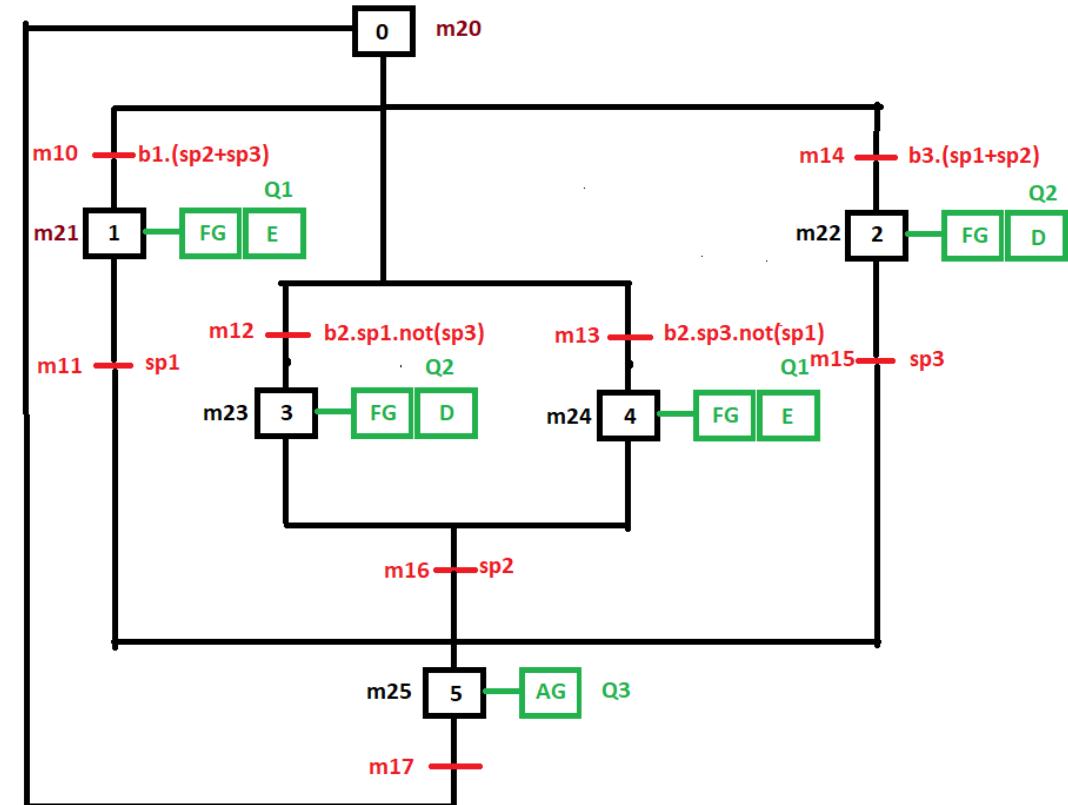
Um sistema de transporte do tipo pórtico (carro que se desloca para Direita (D) ou Esquerda (E) e possui Garra acoplada, a qual pode receber comandos de Abrir (AG) e Fechar (FG)) representado na figura e cujo funcionamento é o seguinte:

- a) o operador seleciona o posto (botões P1, P2 ou P3, do tipo PUSH BUTTON) para onde se deve deslocar o carro; (considere que o carro (ou robô) tem o mapeamento de onde se localiza exatamente cada posto, seja coordenadas no espaço, seja quantos centímetros ou metros precisa se locomover para atingir o destino. Ao chegar no destino há um sensor indicando a presença em P1, P2 ou P3)
- b) este, conforme a sua posição inicial, desloca-se para a esquerda ou para direita;
- c) ao receber o comando para onde deve se deslocar, a garra é fechada. Ao chegar no posto de destino, a garra é aberta
- d) neste sistema, a parti do estado de repouso, o sistema pode executar uma de várias sequências de operações alternativas
- e) depois de concluída essa sequência de operações, o sistema regressa ao seu estado de repouso, que é a última posição solicitada. Ou seja, na condição inicial, pode-se considerar que o carro está em qualquer posição P1, P2 ou P3. Repouso neste caso significa que não há movimento.
- f) Lembre que quando as sequências alternativas são exclusivas (como acontece neste caso), é necessário garantir que as transições de entrada dessas sequências são incompatíveis, isto é, que nunca podem ser simultaneamente verdadeiras, mesmo em caso de avaria ou de erro do operador.

# Problema robô-garra



sketch



Melhorado...

# Problema homegeneizador

Um homegeneizador industrial, mostrado na figura 1, efetua as seguintes operações:

- quando o botão Liga é pressionado, a válvula V1 abre e a matéria-prima, em forma líquida, é inserida no tanque;
- quando o líquido atingir o sensor de nível alto SNA, fecha-se a válvula V1 e inicia-se o processo de homogeneização, **acionando o motor do misturador M1 por 10s;**
- transcorrido esse tempo, abre-se a válvula de saída V2 até que o nível do tanque esteja abaixo do sensor de nível baixo SNB.

Este ciclo deve ser repetido automaticamente por **três vezes**, devendo a operação ser **reiniciada** quando for pressionado novamente o botão liga.

**Obs.: os sensores ficam em nível 1 quando detectam a presença de líquido.**

Como seria o GRAFCET para este problema?

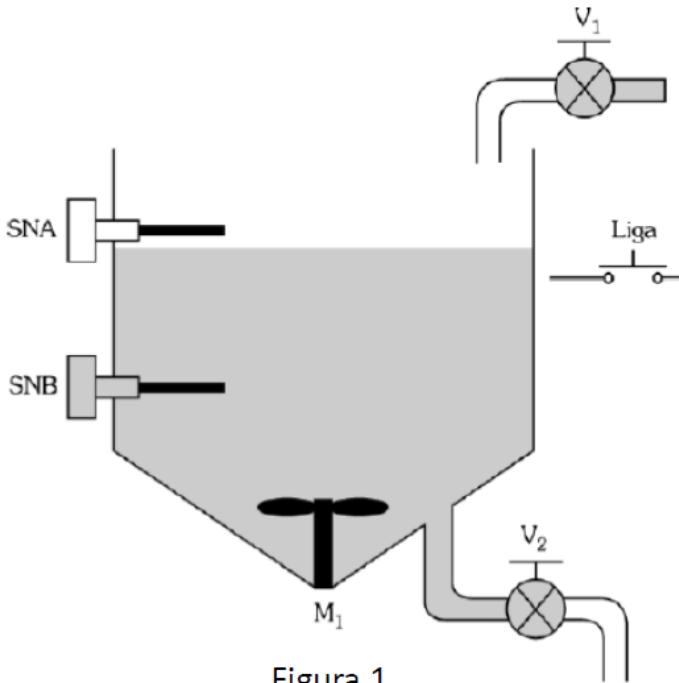


Figura 1

# Casos de uso de modalidades de sistemas embarcados

## Caso 1) software embarcado que controla um eletrodoméstico simples



em paralelo

### Requisitos:

- Interface simples com usuário (botões e display alfanumérico);
- Funcionalidade extremamente bem definida (executar a batida conforme o usuário escolheu, nada mais) ;
- Não possui alta criticidade (se a execução falhar, ninguém se machuca)

### Modalidades elegíveis:

- Bare-metal, pela baixa complexidade exigida do software;
- Ou RTOS, caso seja desejado melhor controle de funcionalidades executadas

<https://embarcados.com.br/prototipacao-plataformas-de-hardware-e-software-para-sistemas-embarcados/>

# Casos de uso de modalidades de sistemas embarcados

**Caso 2) software embarcado que será executado em um sistema crítico de um avião comercial**



## Requisitos:

- Ser extremamente confiável, pois um erro pode significar a morte de muitas pessoas;
- Exige tempo de latência mínimo (um erro na latência pode significar um acidente aéreo);
- Ter gerenciamento de CPU e memória extremamente confiável, resultando em menores chances possíveis de erros de execução.

## Modalidade elegível:

- RTOS, uma vez que possui baixa latência de execução, utiliza um kernel com ótimo gerenciamento de memória e CPU e permite que micro-funcionalidades tenham o menor acoplamento possível.

# Casos de uso de modalidades de sistemas embarcados

**Caso 3) um dispositivo eletrônico em uma loja para colher dados de pesquisa de satisfação do consumidor**

## Requisitos:



- Permitir interação com o usuário, seja por interfaces elaboradas ou mais simples;
- Latência não é algo crítico aqui, afinal demorar alguns milissegundos a mais para executar não atrapalha a vida de ninguém.

## Modalidades elegíveis:

- Sistema operacional de propósito geral, uma vez que latência não é um problema e pode oferecer suporte a conectividades diversas e suporte a interfaces gráficas elaboradas ou não;
- Ou RTOS, caso deseje o menor custo possível com hardware na solução final.

# Casos de uso de modalidades de sistemas embarcados

Caso 4) uma smart TV, com suporte a conectividade Bluetooth e WiFi e a aplicativos de streaming muito populares do mercado



## Requisitos:

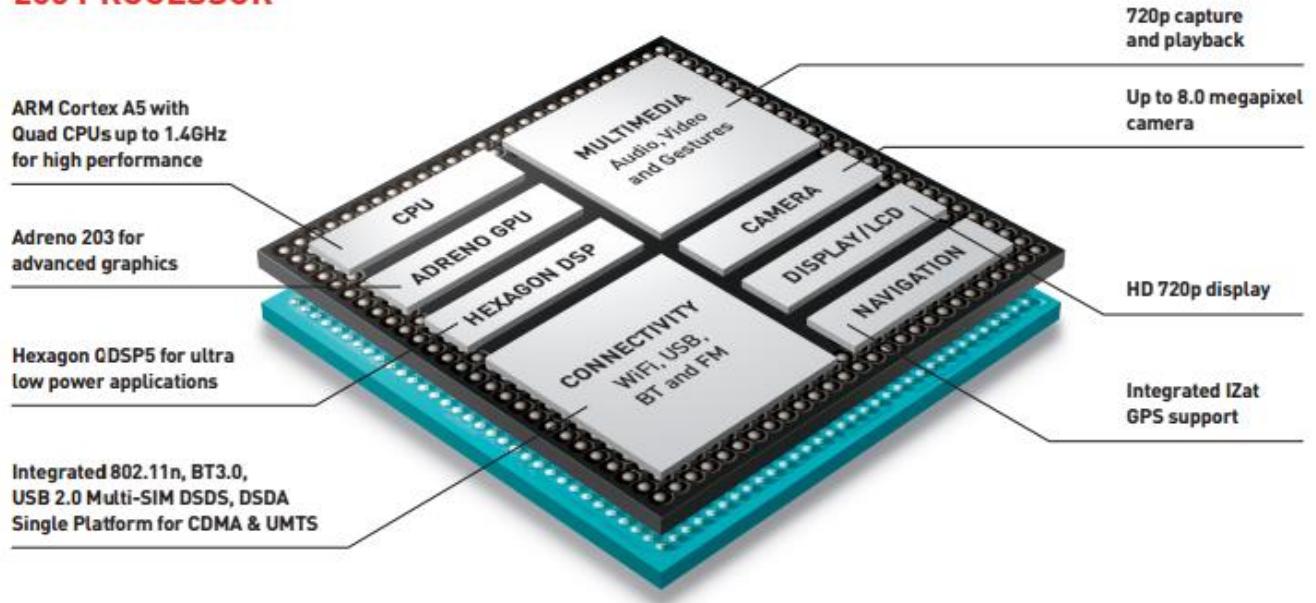
- Interfaces com o usuário devem ser, obrigatoriamente, muito elaboradas;
- Suportar diversos tipos de conectividade de forma simultânea;
- Suportar aplicações terceiras;
- Latência não é algo crítico aqui, afinal demorar alguns milissegundos a mais para executar não atrapalha a vida de ninguém

## Modalidade elegíveis:

- Sistema operacional de propósito geral somente.

# Arquitetura ARM e System on a Chip (SoC)

## 200 PROCESSOR



Cortex A73/A53



- Integração de componentes num único chip (subsistemas) – CPU, memória, E/S, tratamento de radiofrequência, gerenciamento de bateria, sensores
- Dispositivos menores, com menor consumo de energia
- Ideal para computação móvel, embarcada (ou embutida, **embedded systems**)
- Smartphones, tablets, câmeras digitais, e-book reader, consoles, gps, microcontroladores



# Arquitetura ARM e desenvolvimento de software

Download the Arduino IDE

The screenshot shows the Arduino IDE download page. At the top right, it says "2.1.1 em Agosto23". Below that, the version "ARDUINO 1.8.13" is displayed. To the left is the Arduino logo, which is a teal circle containing a white infinity symbol with a minus sign on the left and a plus sign on the right. The main content area has a teal background. It lists download links for different operating systems:

- Windows** Installer, for Windows 7 or newer
- Windows** ZIP file for non-Windows users
- Windows app** Requires Windows 10 or newer
- Mac OS X** 10.10 or newer
- Linux** 32 bits
- Linux** 64 bits
- Linux** ARM 32 bits
- Linux** ARM 64 bits

Links at the bottom include "Release Notes", "Source Code", and "Checksums (sha512)". A red oval highlights the "Linux" download options.

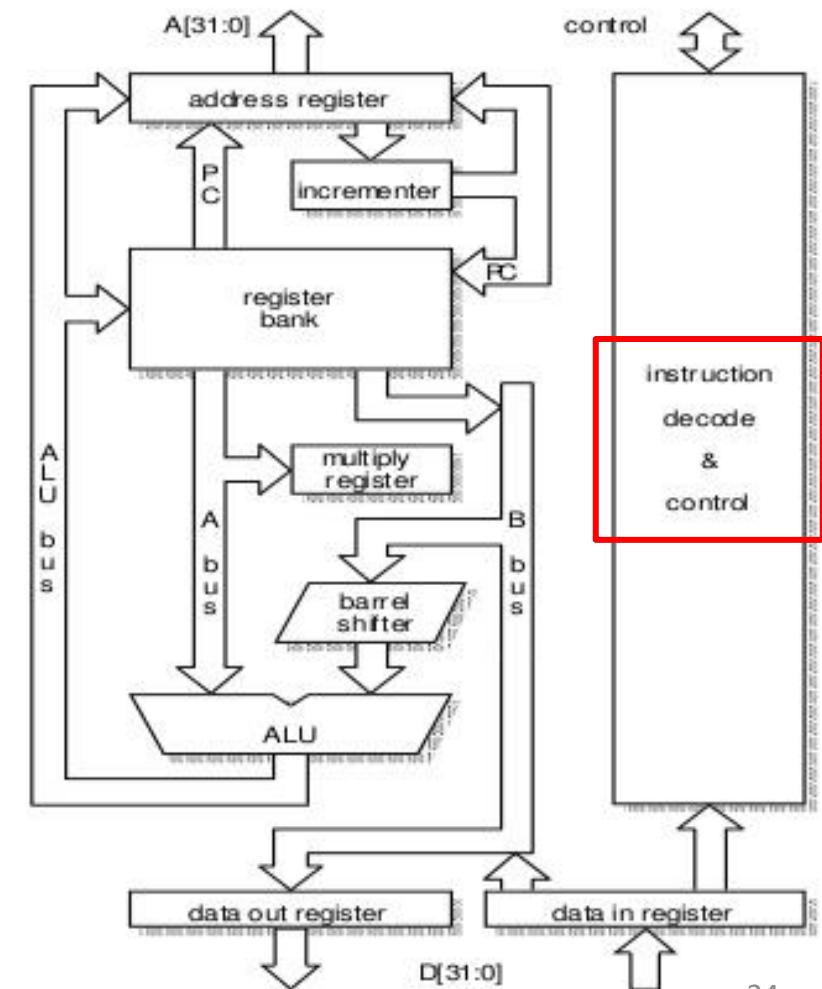


Raspberry Pi 4 Model B  
**Broadcom 2711**  
Quad-core Cortex-A72  
(ARMv8-A) 64-bit SoC  
@ 1.5 GHz. X 4 threads

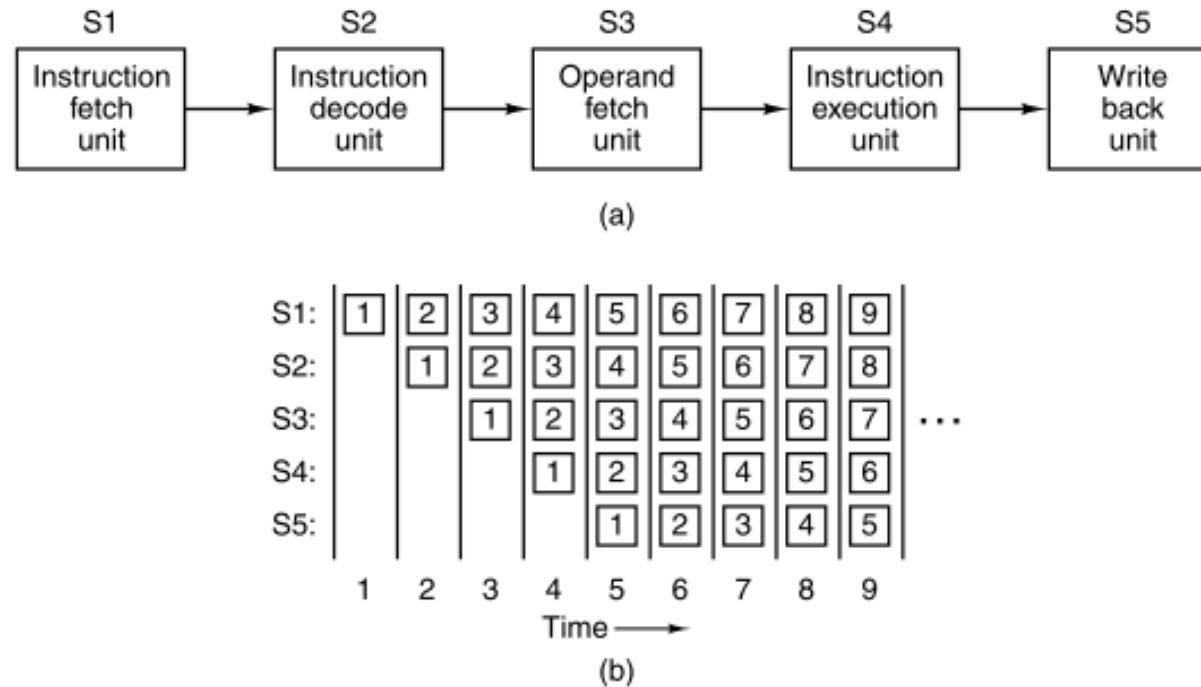
# Arquitetura ARM (ARM Holdings, Cambridge, Inglaterra, arm.com)

- Família de arquiteturas **RISC** para processadores
  - Set de instruções menor, instruções mais simples
  - Acesso simples à memória, menos ciclos para obter operandos
  - Set de registradores (register bank) para permitir **pipeline**, várias instruções executadas ao mesmo tempo
  - Programas maiores (assembler) ocupam mais memória
- Licenciada para outros fabricantes, podem criar seus próprios Chips, mas usam esta arquitetura como base (Samsung, Apple, Nvidia, Qualcomm, Texas Instruments etc.)
- **37 registradores** (30 de propósito geral)
- Instruções  $\geq 16$  bits (set de instruções thumb)
- Variantes (extensões):
  - NEON (áudio, vídeo), VFP (autotronica, gráfico, 3D, indústria)
  - DSP (sinais digitais), Jazelle (java), big.Little (múltiplos processadores) – no momento dynamIQ
  - TrustZone (proteção e segurança digital)

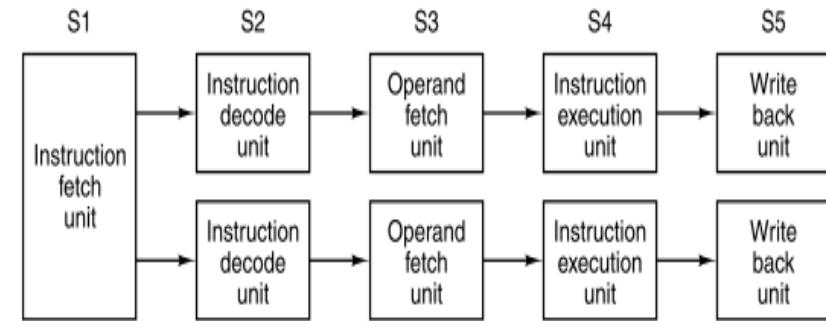
## The ARM Architecture



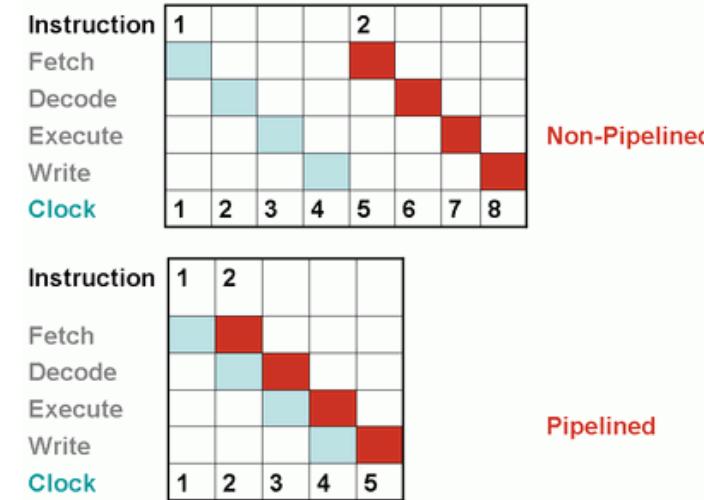
# Noções sobre multiprocessamento (múltipla execução simultânea)



**Figure 2-4.** (a) A five-stage pipeline. (b) The state of each stage as a function of time. Nine clock cycles are illustrated.

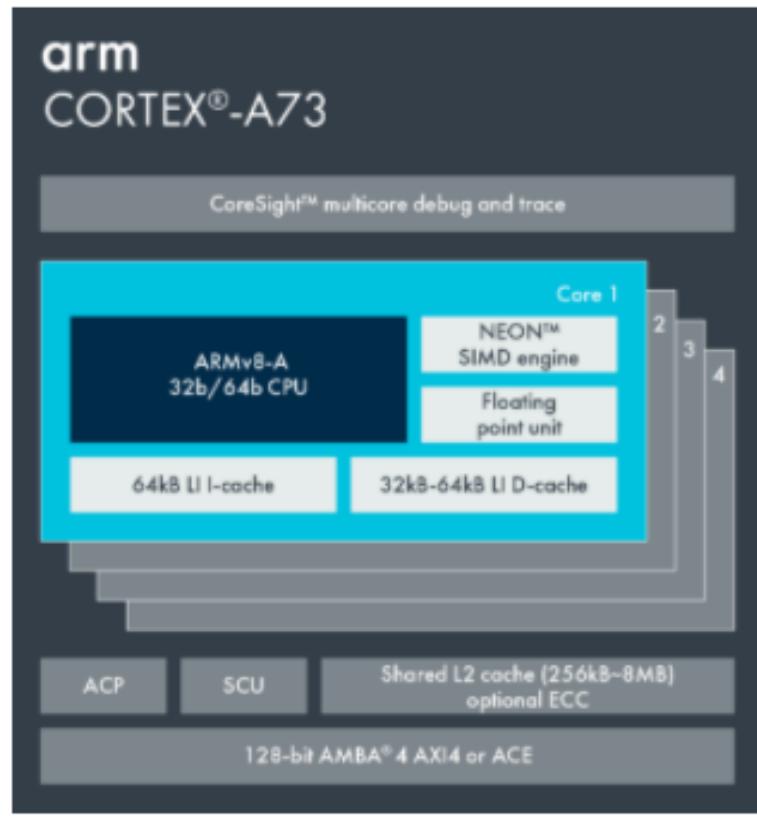


**Figure 2-5.** Dual five-stage pipelines with a common instruction fetch unit.



Pipeline em unidade individual de processamento  
Superescalar: múltiplas unidades de processamento

# Arquitetura ARM e desenvolvimento de software



Arm Cortex-A73 CPU

Architecture	Armv8-A <b>Desde 2011</b>
Multicore	1-4x Symmetrical Multiprocessing (SMP) within a single processor cluster, and multiple coherent SMP processor clusters through AMBA 4 ACE technology
ISA Support	<ul style="list-style-type: none"><li>AArch32 for full backward compatibility with Armv7</li><li>AArch64 for 64-bit support and new architectural features</li><li><u>TrustZone</u> security technology</li><li><u>Neon</u> advanced SIMD</li><li>DSP &amp; SIMD extensions</li><li>VFPv4 floating point</li><li>Hardware virtualization support</li></ul>
Debug & Trace	<u>CoreSight SoC-400</u>

<https://developer.arm.com/ip-products/processors/cortex-a/cortex-a73>

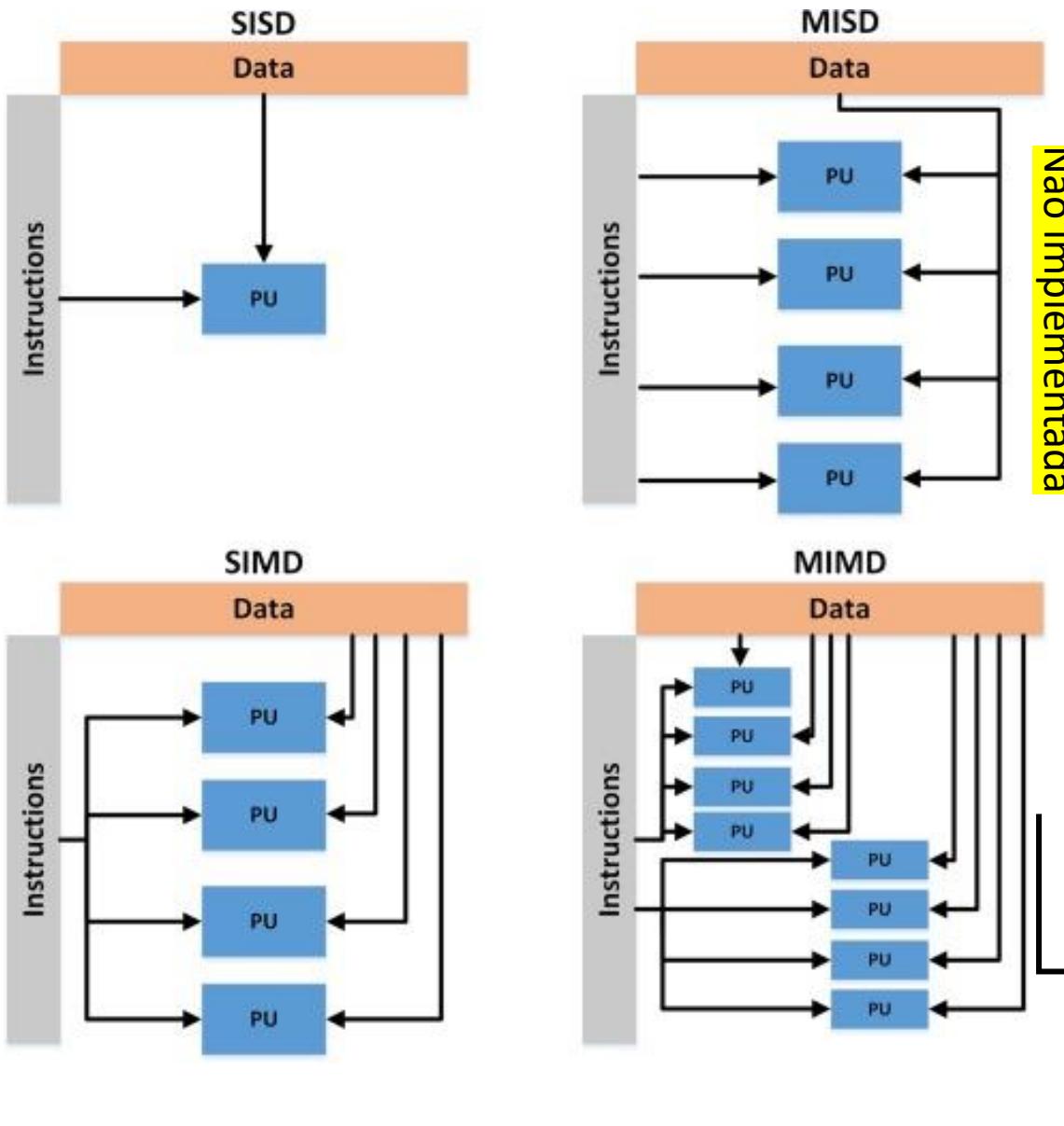
**Em 2021 anunciou Armv9**

## Latest Versions of the Arm Architecture

Arm works with its partners to evolve its architecture and meet future needs. Armv9-A is a set of extensions to the Armv8-A architecture, and part of a rolling program of substantial enhancements to the architecture to be deployed over the next few years. The goal of these enhancements is to help increase the computing capability in areas such as digital signal processing (DSP) and machine learning (ML), and to continually improve the security and robustness of our systems.

The latest architectures for the R-profile and M-profile are Armv8-R and Armv8-M.

- Noções multiprocessamento (arquiteturas) – classificação de Flynn



```

import java.util.Arrays;
import java.util.List;

public class TestStream {
    public static void main(String[] args) {
        // stream sequencial
        List<Integer> listOfNumbers = Arrays.asList(1, 2, 3, 4, 10, 20, 30, 40);
        listOfNumbers
            .stream() //stream API (Collections) desde 2014, Java 8
            .forEach(number ->
                System.out.println(number + " " + Thread.currentThread().getName())
            );
        System.out.println("Regiao Paralela: \n");
        // stream em paralelo
        listOfNumbers
            .parallelStream()
            .forEach(number ->
                System.out.println(number + " " + Thread.currentThread().getName())
            );

        // soma em paralelo e reduz para sum, adicionando 5 à soma total (0 valor inicial)
        int sum = listOfNumbers
            .parallelStream()
            .reduce(0, Integer::sum) + 5; // lambda: (subtotal, element) -> subtotal + element)
        System.out.println(sum);
    }
}

```

SISD

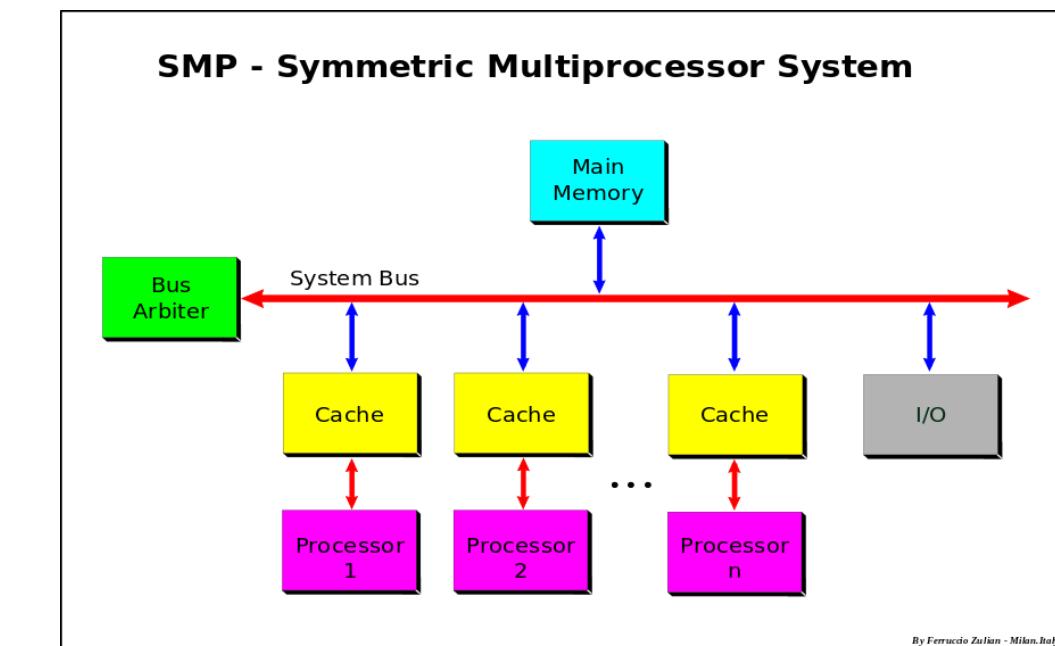
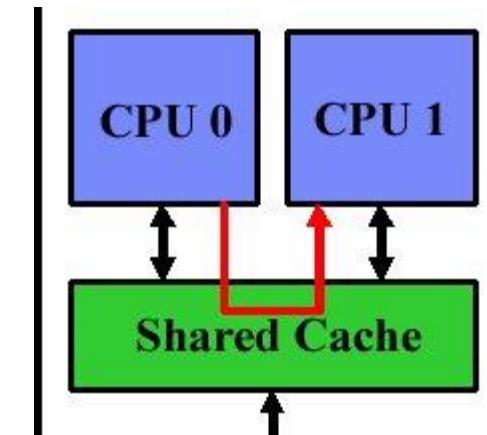
SIMD

## • Noções multiprocessamento

Processadores de uso geral

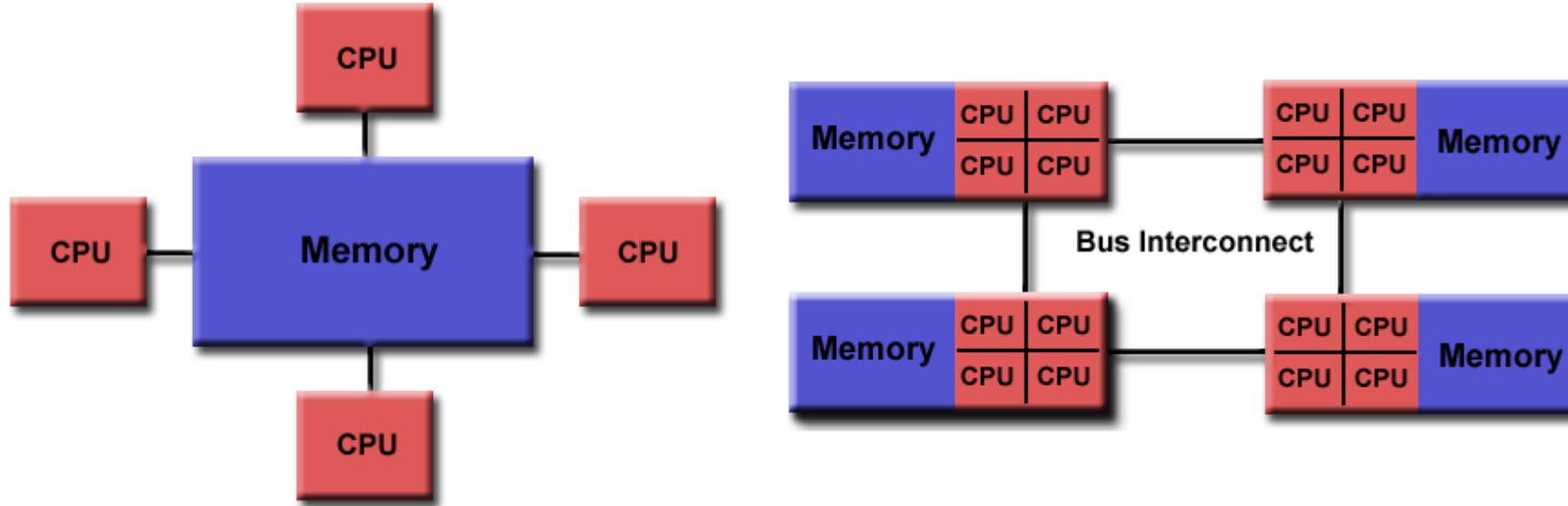
**Solução 1:** memória compartilhada: cada processador acessa programas e dados armazenados na memória compartilhada e os processadores se comunicam uns com os outros por meio dessa memória SMP – multiprocessador simétrico – múltiplos processadores compartilham uma única memória ou um pool de memória por um barramento compartilhado; o tempo de acesso é aproximadamente o mesmo para cada processador (**UMA**), embora na arquitetura variante **NUMA** o tempo de acesso possa ser diferenciado (não uniforme).

**Solução 2:** memória distribuída:

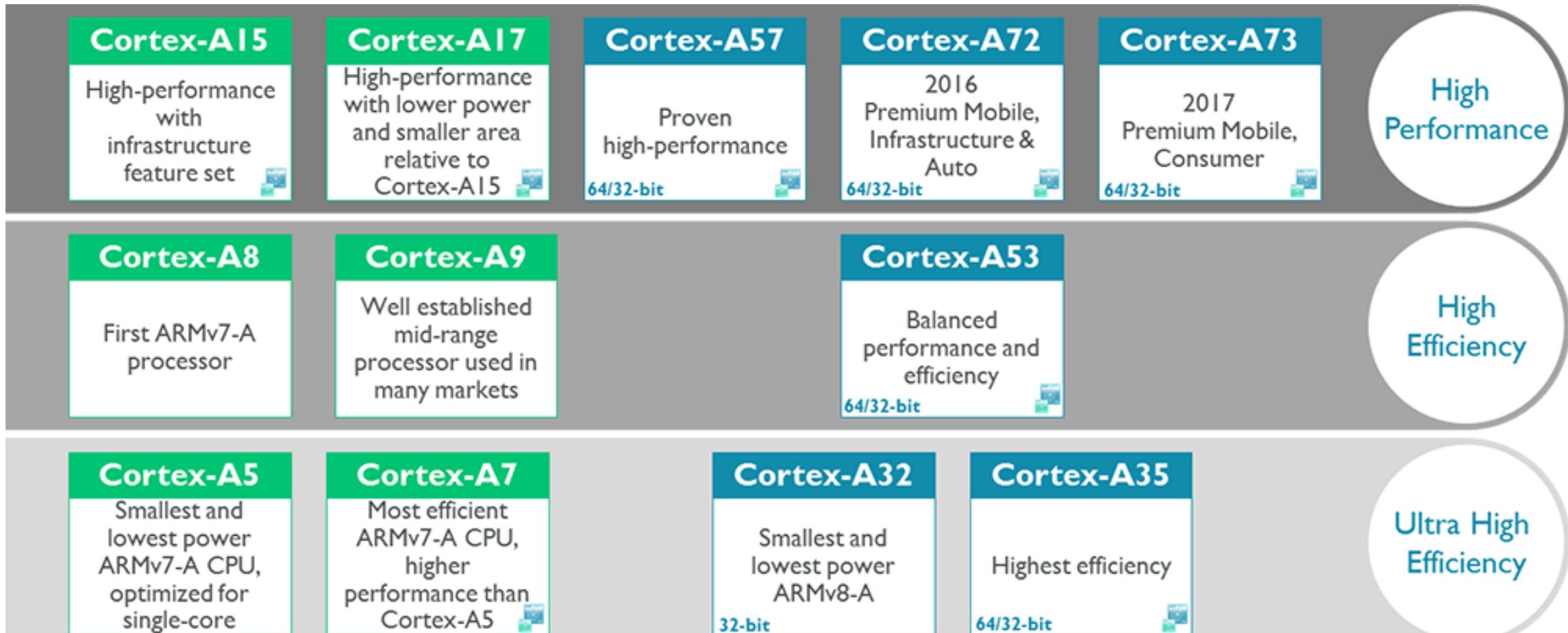


By Ferruccio Zulian - Mikon.Baby

- Noções multiprocessamento com memória compartilhada



# Arquitetura ARM e desenvolvimento de software



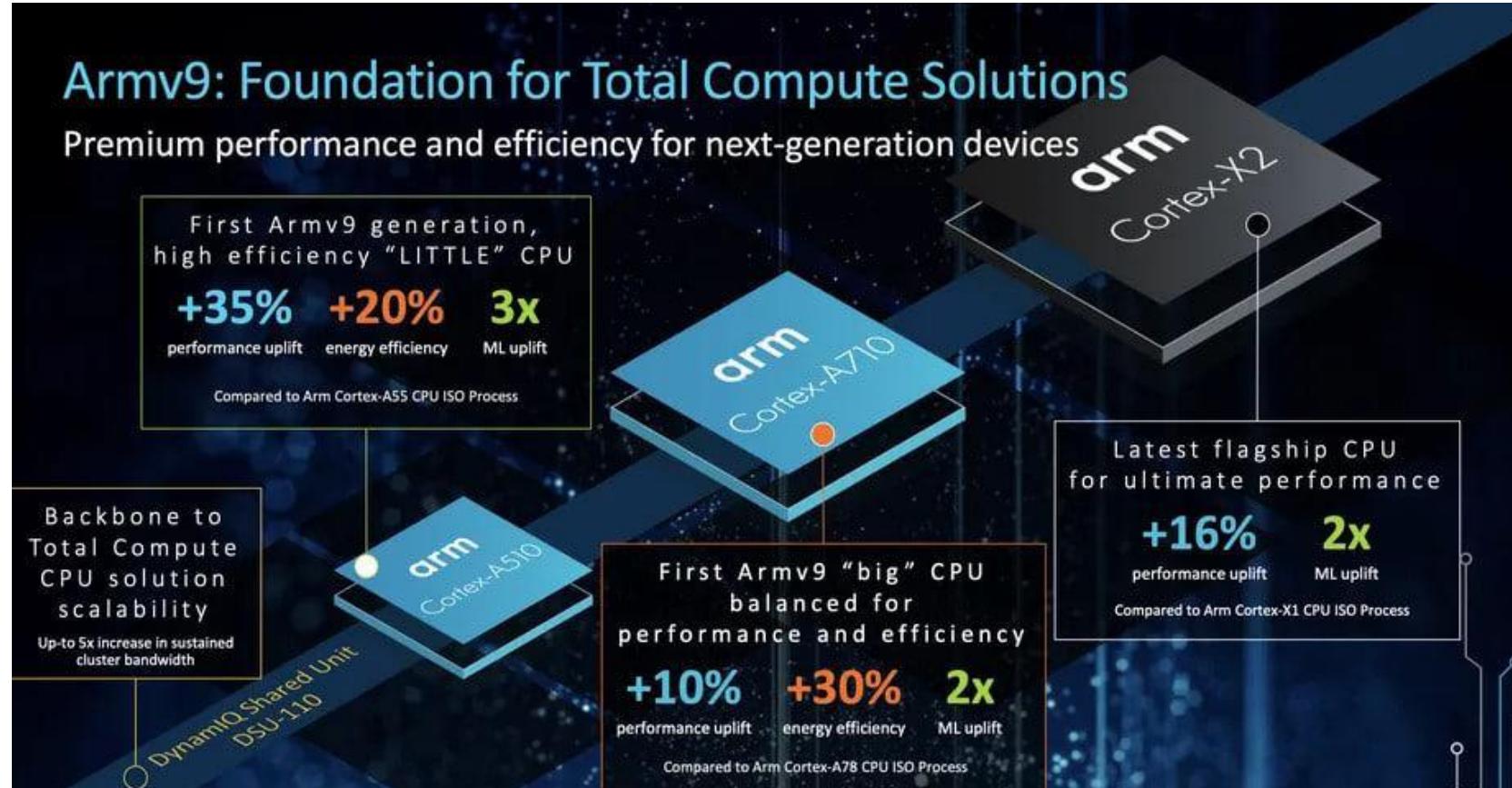
ARMv7-A

ARMv8-A

Key: big.LITTLE compatible

# A ideia do big.LITTLE

há uma combinação de um núcleo mais poderoso para tarefas que exigem “força bruta”, com núcleos menores e mais eficientes para tarefas mais leves.



Exemplo: Snapdragon 888  
o conjunto é formado por:

- um Cortex-X1,
- três Cortex-A78 (big)
- quatro Cortex-A55 (LITTLE),
- GPU Adreno

Fonte: <https://olhardigital.com.br/2021/05/25/reviews/armv9-arm-apresenta-as-primeiras-cpus-e-gpus-baseadas-na-nova-arquitetura/>

# GPUs com foco em ML (machine learning)



Fonte: <https://olhardigital.com.br/2021/05/25/reviews/armv9-arm-apresenta-as-primeiras-cpus-e-gpus-baseadas-na-nova-arquitetura/>

# Uma rede neural é essencialmente paralela!

AI everywhere demands specialized, scalable solutions

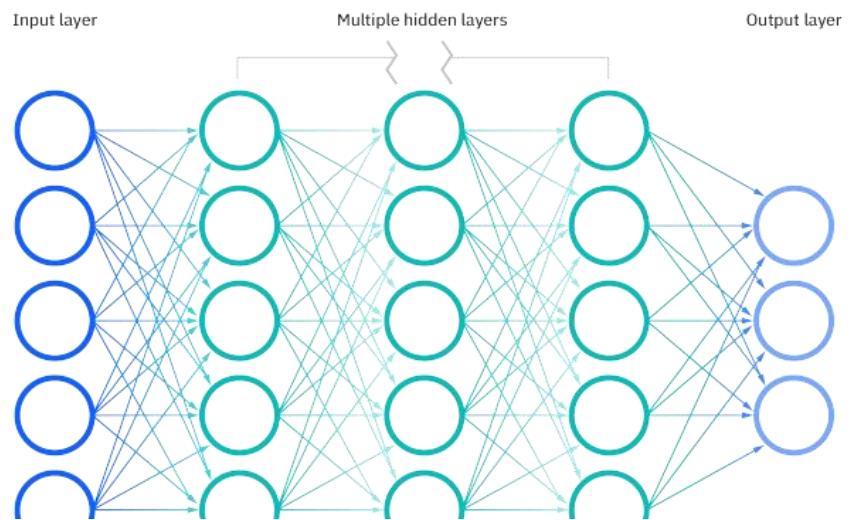
The ubiquity and range of AI workloads demands more diverse and specialized solutions. For example, it is estimated there will be more than eight billion AI-enabled voice-assisted devices in use by the mid-2020s<sup>i</sup>, and 90 percent or more of on-device applications will contain AI elements along with AI-based interfaces like vision or voice<sup>ii</sup>.

To address this need, Arm partnered with Fujitsu to create the Scalable Vector Extension (SVE) technology, which is at the heart of Fugaku, the world's fastest supercomputer. Building on that work, Arm has developed SVE2 for Armv9 to enable enhanced machine learning (ML) and digital signal processing (DSP) capabilities across a wider range of applications.

SVE2 enhances the processing ability of 5G systems, virtual and augmented reality, and ML workloads running locally on CPUs, such as image processing and smart home applications. Over the next few years, Arm will further extend the AI capabilities of its technology with substantial enhancements in matrix multiplication within the CPU, in addition to ongoing AI innovations in its Mali™ GPUs and Ethos™ NPUs.

<https://www.arm.com/company/news/2021/03/arms-answer-to-the-future-of-ai-armv9-architecture>

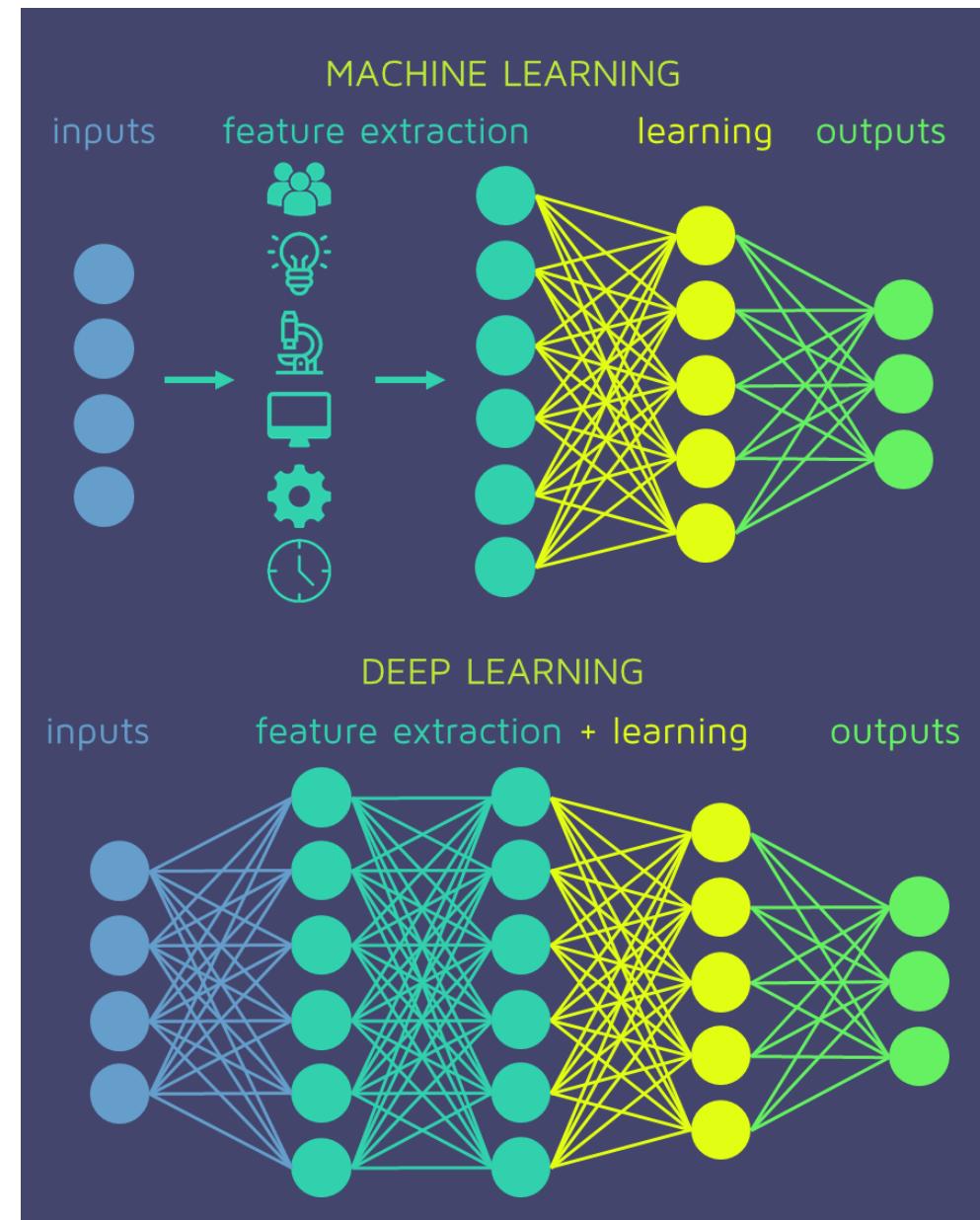
# Uma rede neural é essencialmente paralela!



Para saber mais:

- Redes Neurais Convolucionais (CNN): visão
- Redes Neurais Recorrentes/Transformers: PLN
- tinyML

<https://conect2ai.dca.ufrn.br/>



# Uma rede neural é essencialmente paralela!

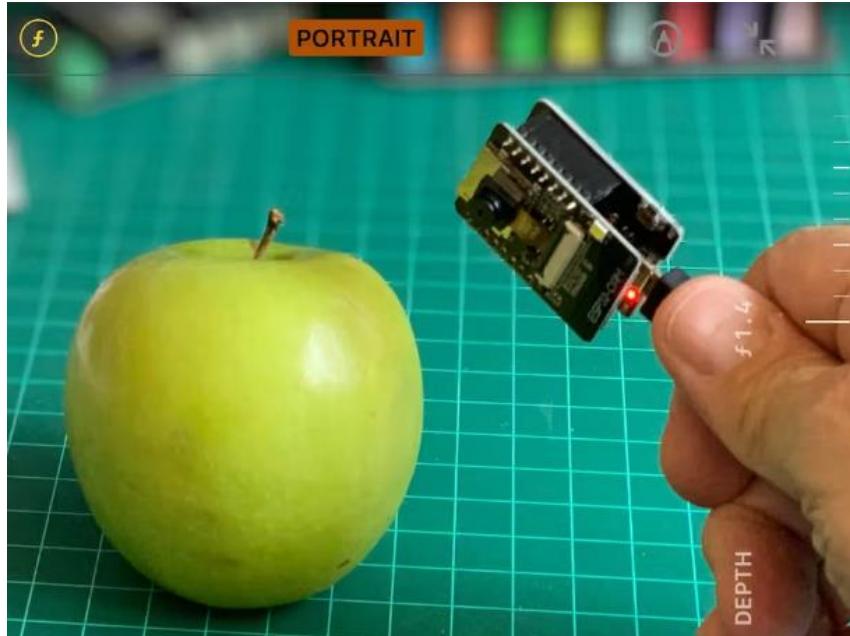
## Como funciona uma Alexa Skill

Uma skill de Alexa tem tanto um modelo de interação - ou interface de usuário de voz - como uma lógica de aplicativo. Quando um cliente fala, a Alexa processa a fala no contexto do seu modelo de interação para determinar qual foi a solicitação do cliente. A Alexa então envia a solicitação à sua lógica de aplicação de skills, que processa essa informação. Numa skill, você fornece a lógica do aplicativo como um serviço de nuvem de back-end hospedado pela Alexa, AWS ou outro servidor.



<https://developer.amazon.com/pt-BR/alexza/alex-skills-kit/start>

# Mas e com processadores mais “tímidos”?



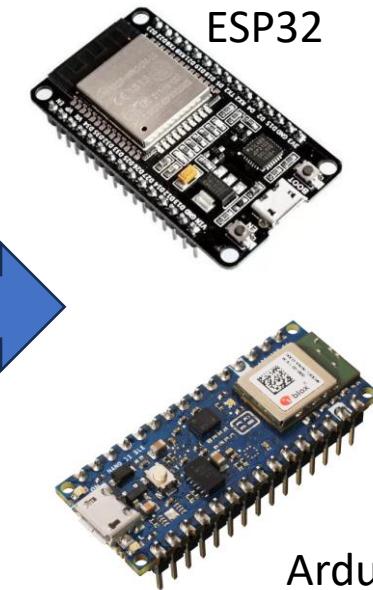
EDGE IMPULSE

TensorFlow  
Lite

Treino dos modelos



Deploy



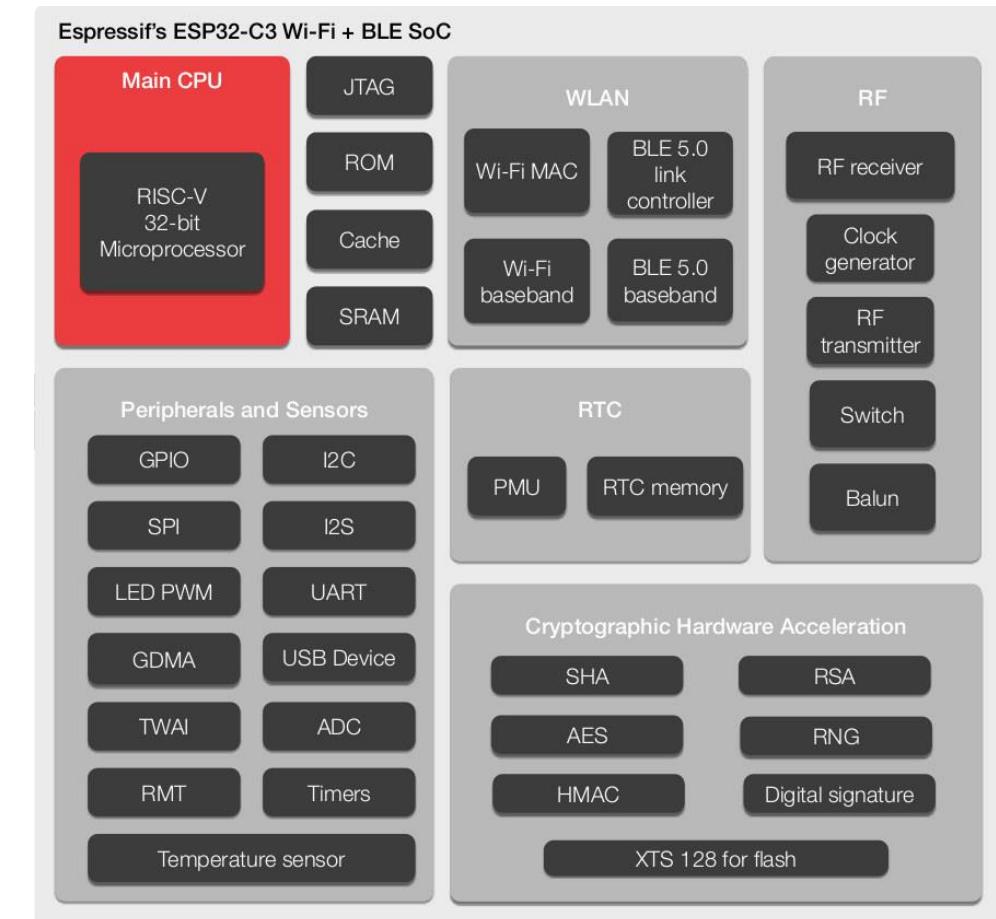
Arduino BLE Nano 33

<https://www.hackster.io/mjrobot/esp32-cam-tinyml-image-classification-fruits-vs-veggies-4ab970>

# Mas e com processadores mais “tímidos”?

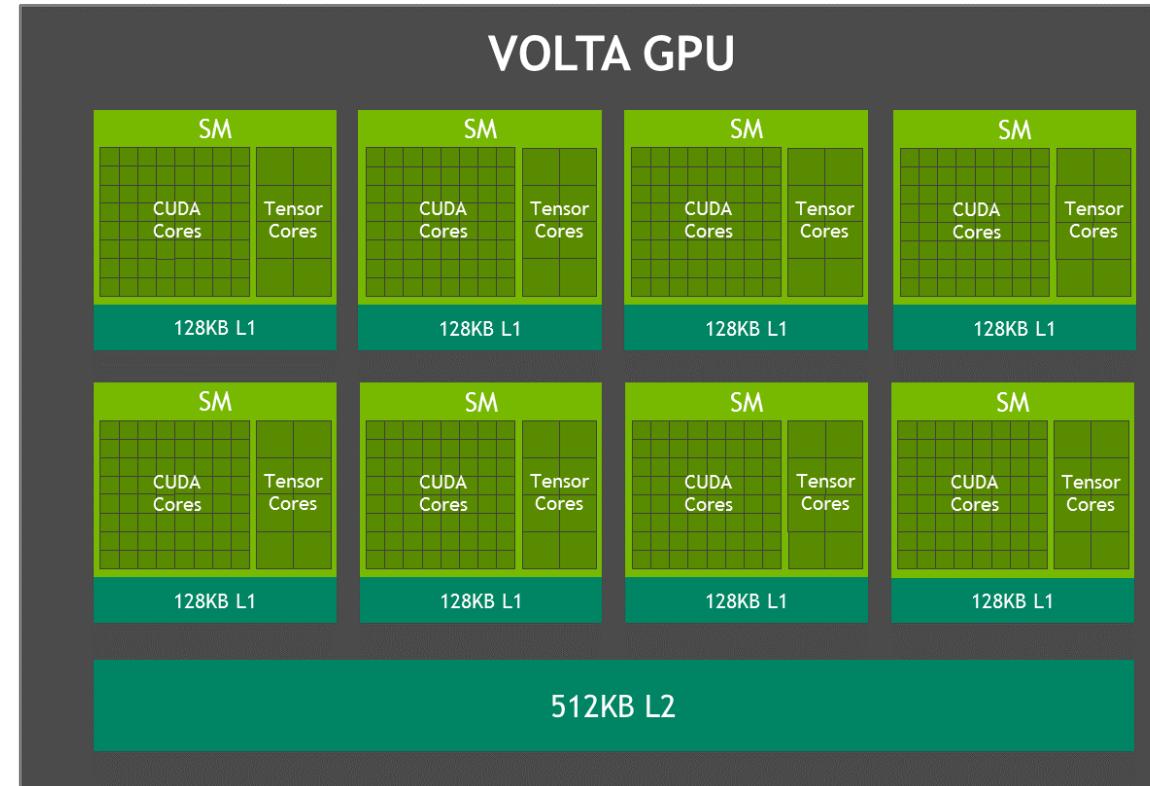
Specifications	ESP32
<b>MCU</b>	Xtensa Dual-Core 32-bit LX6 600 DMIPS
<b>802.11 b/g/n Wi-Fi</b>	Yes, HT40
<b>Bluetooth</b>	Bluetooth 4.2 and below
<b>Typical Frequency</b>	160 MHz
<b>SRAM</b>	512 kBytes
<b>Flash</b>	SPI
<b>GPIO</b>	36
<b>Hardware / Software PWM</b>	1 / 16 Channels
<b>SPI / I2C / I2S / UART</b>	4/2/2/2
<b>ADC</b>	12-bit
<b>CAN</b>	1
<b>Ethernet MAC Interface</b>	1
<b>Touch Sensor</b>	Yes
<b>Temperature Sensor</b>	Yes
<b>Working Temperature</b>	-40° C – 125° C

- Flash Memory: 4 MB (8/16) – módulos externos...
- SRAM: 520 KB
- Clock Speed: 240 Mhz



# Plataformas com foco em IA/ML (nvidia)

[https://developer.nvidia.com/embedded/community/jetson-projects?sortBy=jetson\\_community\\_projects%2Fsort%2Fdate\\_added%3Adesc](https://developer.nvidia.com/embedded/community/jetson-projects?sortBy=jetson_community_projects%2Fsort%2Fdate_added%3Adesc)



Versão Orin Nano: US\$ 560

<https://developer.nvidia.com/blog/nvidia-jetson-agx-xavier-32-teraops-ai-robotics/>

# Plataformas com foco em IA/ML (nvidia)

## Stencil computations on GPU

Using the `numba.cuda` module I'm able to get about a 200x increase with a modest increase in code complexity.

```
In [3]: from numba import cuda

@cuda.jit
def smooth_gpu(x, out):
    i, j = cuda.grid(2)
    n, m = x.shape
    if 1 <= i < n - 1 and 1 <= j < m - 1:
        out[i, j] = (x[i - 1, j - 1] + x[i - 1, j] + x[i - 1, j + 1] +
                      x[i      , j - 1] + x[i      , j] + x[i      , j + 1] +
                      x[i + 1, j - 1] + x[i + 1, j] + x[i + 1, j + 1]) // 9
```

```
In [4]: import cupy, math

x_gpu = cupy.ones((10000, 10000), dtype='int8')
out_gpu = cupy.zeros((10000, 10000), dtype='int8')

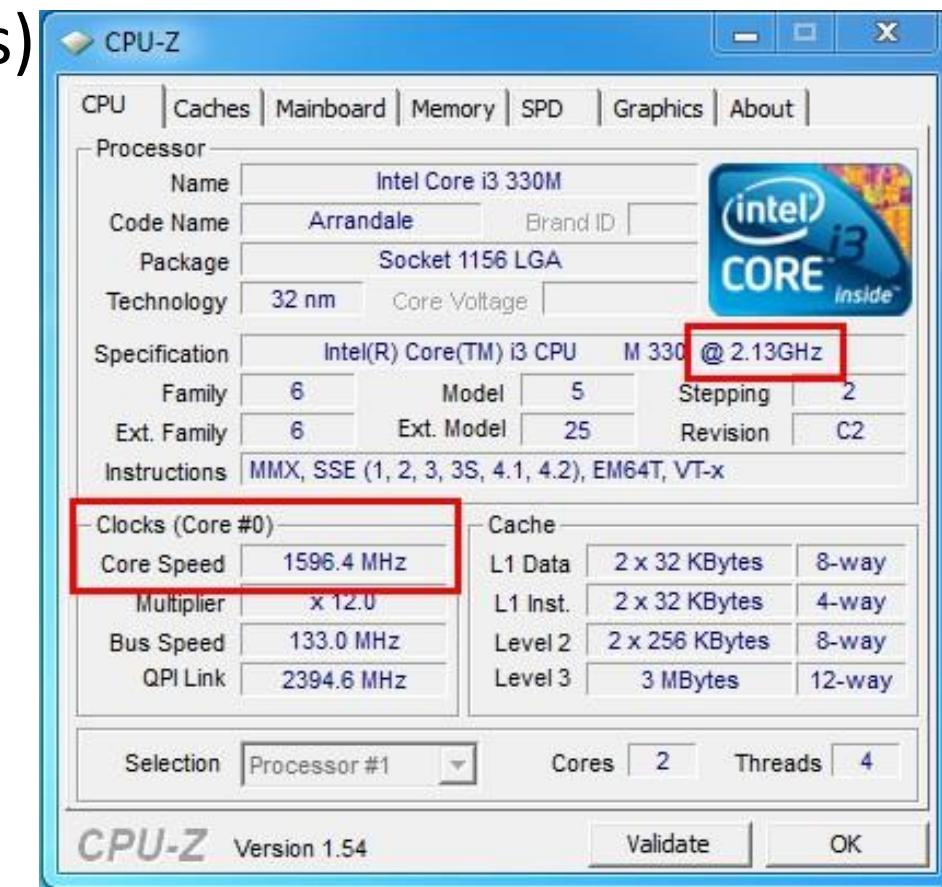
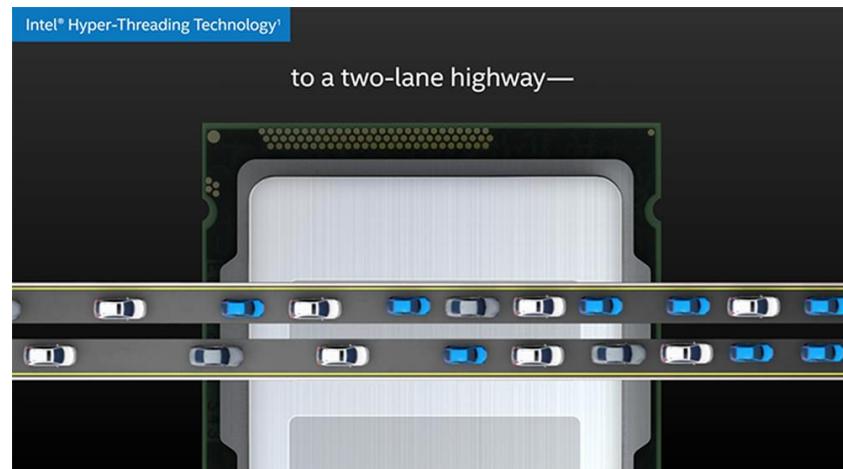
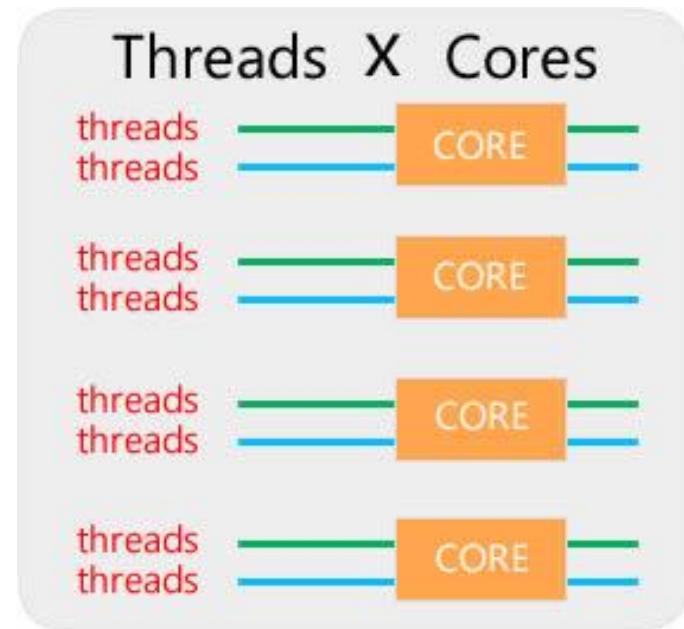
# I copied the four lines below from the Numba docs
threadsperblock = (16, 16)
blockspergrid_x = math.ceil(x_gpu.shape[0] / threadsperblock[0])
blockspergrid_y = math.ceil(x_gpu.shape[1] / threadsperblock[1])
blockspergrid = (blockspergrid_x, blockspergrid_y)

%timeit smooth_gpu[blockspergrid, threadsperblock](x_gpu, out_gpu)
```

2.87 ms ± 90.8 µs per loop (mean ± std. dev. of 7 runs, 1000 loops each)

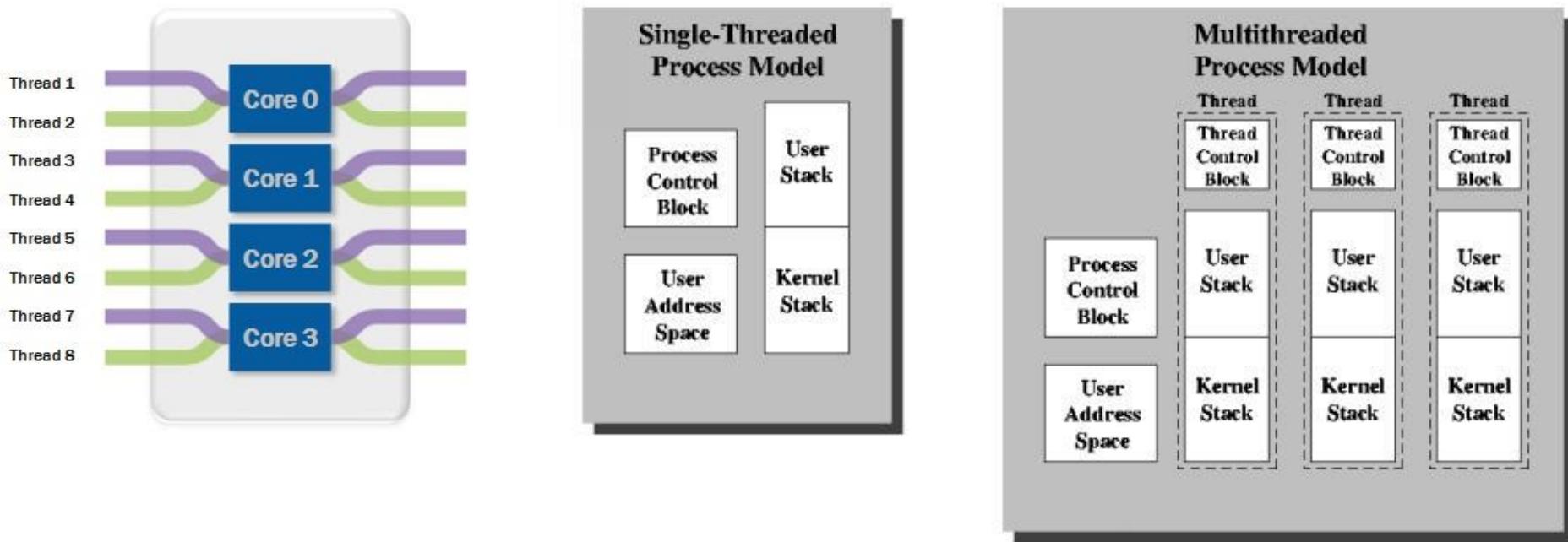


- Noções multiprocessamento (threads)



- Execução simultânea de sequências diferentes de instruções
  - Cada núcleo é uma linha de execução, mas a aplicação precisa ser programada para explorar isto (multi-thread, hyper-thread INTEL)
  - Cada núcleo processando N threads (fluxos paralelos de código)
- Programação paralela C/C++: <pthread.h>, OpenMP, MPI

- Noções multiprocessamento (threads)



*Cada thread com sua própria pilha (Stack)*

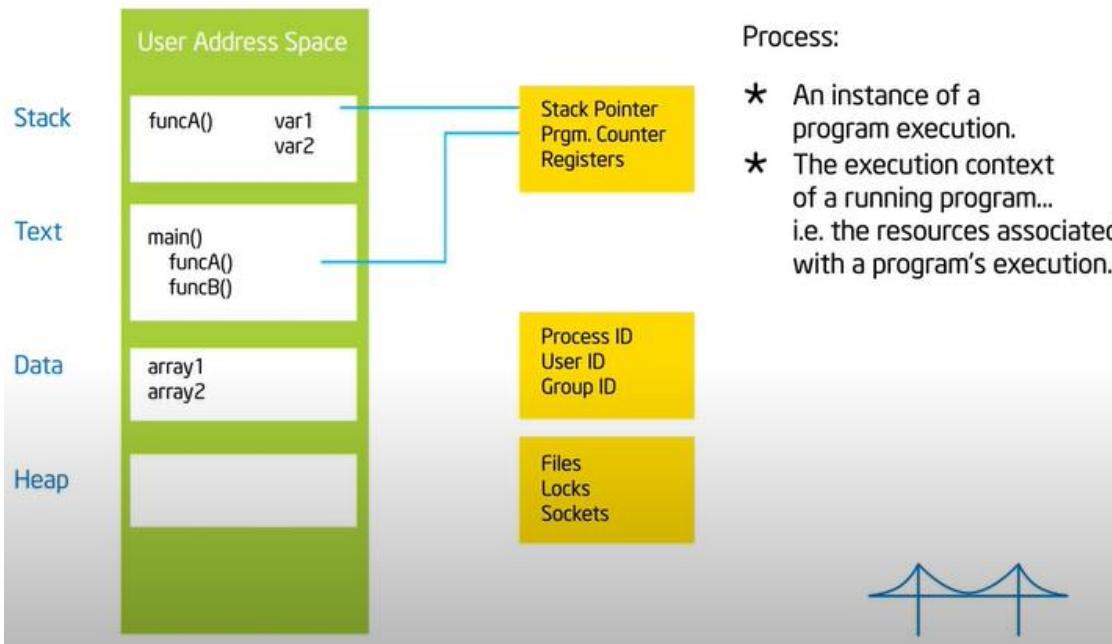
A thread (ou thread de execução) é uma unidade de paralelismo

Possui o necessário para executar um fluxo de instruções – pilha privada, contador de programa, trecho de código  
São interessantes para cooperar em dados globais compartilhados, compartilham também sistema de arquivo

No caso de multiprocessador com cache separada, compartilhando só principal: **coerência de cache!**

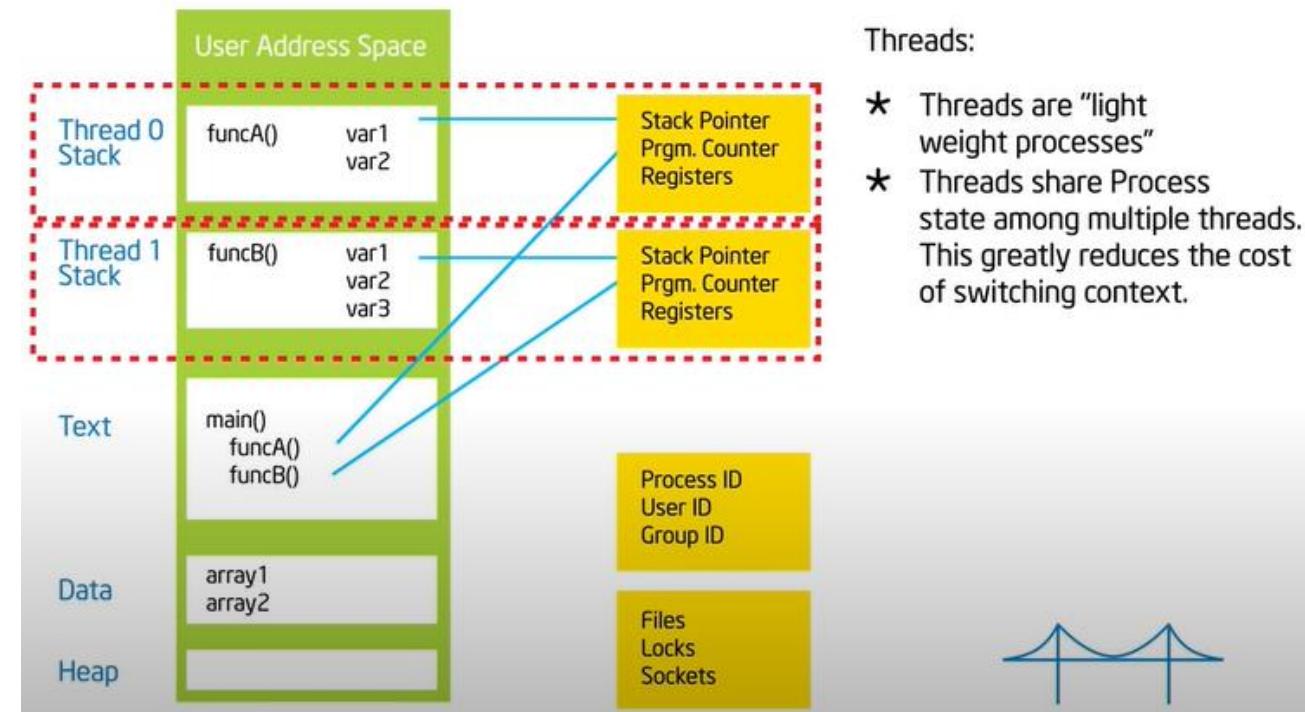
- Noções multiprocessamento (threads)

## Programming Shared Memory Computers



### Process:

- ★ An instance of a program execution.
- ★ The execution context of a running program... i.e. the resources associated with a program's execution.



### Threads:

- ★ Threads are "light weight processes"
- ★ Threads share Process state among multiple threads. This greatly reduces the cost of switching context.

Fonte: Introduction to OpenMP (Tim Mattson – INTEL)

<https://youtu.be/x0HkbIuJILk?si=gauRBIUyCifuifE->

## • Noções multiprocessamento (threads)

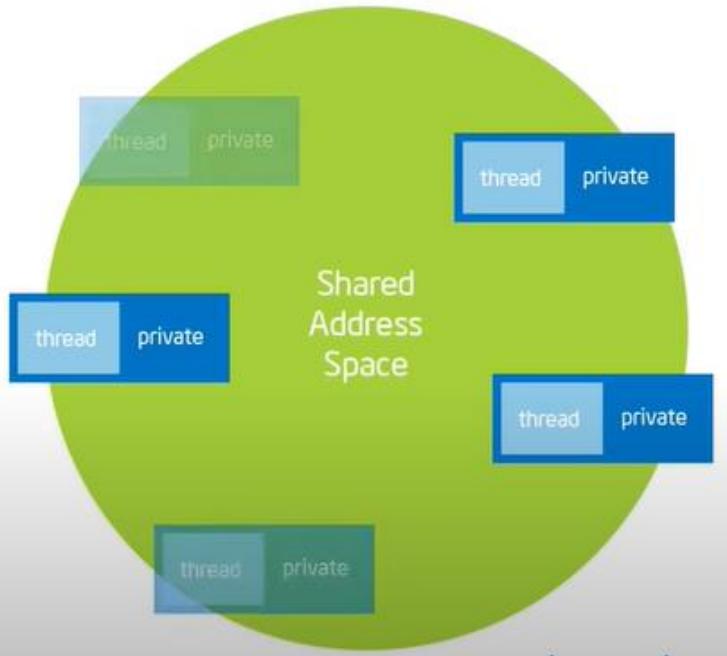
An instance of a program:

One Process and lots of threads

Threads interact through reads/writes to a shared address space

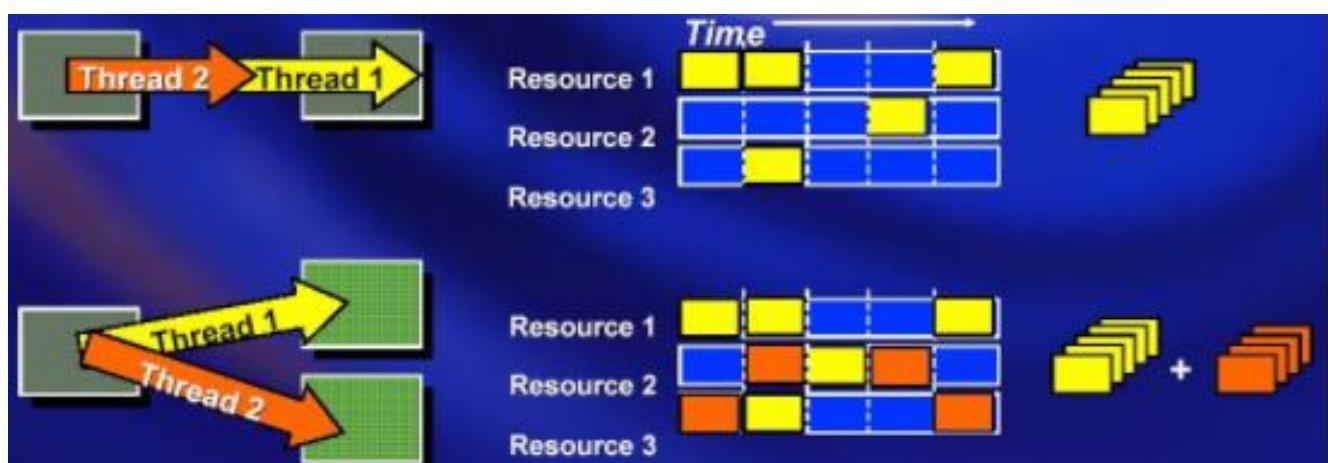
OS scheduler decides when to run which threads... interleaved for fairness

Synchronization to assure every legal order results in correct results



Fonte: Introduction to OpenMP (Tim Mattson – INTEL)

<https://youtu.be/x0HkbluJILk?si=gauRBIUyCifuifE->



- Noções multiprocessamento (threads)

Processador executando partes diferentes de um mesmo programa, ao mesmo tempo

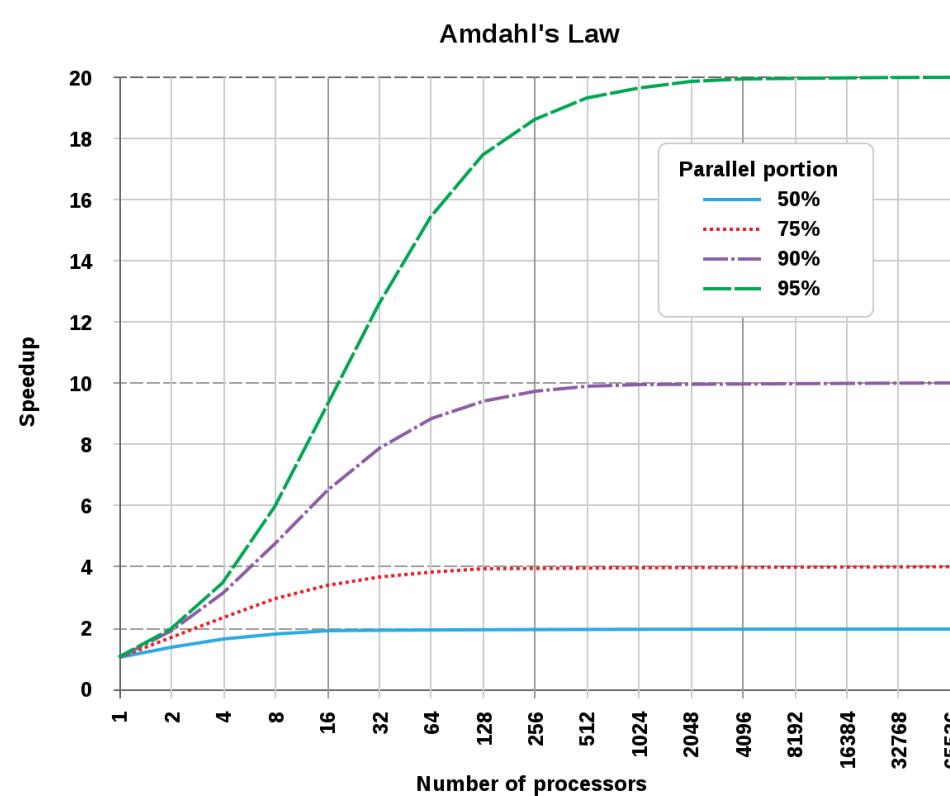
Para o multiprocessamento ser possível

- Os circuitos de apoio da CPU (chipset) devem suportar
- O sistema operacional deve ser capaz de utilizar vários processadores simultaneamente
- O próprio processador deve poder ser usado num sistema multiprocessado
- **Os programas devem ser projetados para tirar proveito do multiprocessamento**

A carga precisa ser distribuída entre os processadores (load balance)

*Automaticamente ou manualmente*

**Uso mais eficiente da cache:** é improvável que uma única thread de execução possa usar efetivamente toda a memória. Um número de threads ou processos relativamente independentes tem uma oportunidade maior de tirar vantagem da memória cache



O speedup de um programa usando múltiplos processadores em computação paralela é limitado pela fração sequencial do programa. Por exemplo, se 95% do programa pode ser paralelo, teoricamente o speedup máximo usando computação paralela seria 20x como apresentado no diagrama, não importando quantos processadores estão sendo usados.

# Multicore

- Desafio para o desenvolvedor: identificar partes do código ‘sequencial’ que podem ser ‘paralelizados’

$$\begin{aligned}\text{Aumento de velocidade (speedup)} &= \frac{\text{tempo para executar o programa em single core}}{\text{tempo para executar o programa em } N \text{ processadores}} \\ &= \frac{1}{(1 - f) + \frac{f}{N \text{ (cores)}}}\end{aligned}$$

Lei de Amdahl: a lei supõe um programa no qual uma fração  $(1-f)$  do tempo de execução envolve código inherentemente serial e uma fração  $f$  que envolve código infinitamente paralelizável

Contudo, numa abordagem direta, não significa aumento de desempenho significativo, pois implica no **aumento de sobrecarga na comunicação e distribuição de trabalho entre vários processadores e sobrecarga de coerência de cache**. A curva de melhoria apresenta picos, mas depois passa a degradar. Por exemplo, se  $f = 0,9$  (apenas 10% de serial), com  $N = 8$ , a melhoria é de 4,7. Contudo melhorias para reduzir a fração serial foram incorporadas ao longo do tempo e *sistemas operacionais, bancos de dados e software para servidores usam intensamente a organização paralela*

## Exercício 1

- Suponha que uma tarefa faz uso intensivo de operações de ponto flutuante, com 40% do tempo consumido por operações de ponto flutuante. Com um novo design de hardware, o módulo de ponto flutuante é acelerado por um fator de K. Calcule o speedup máximo possível, considerando a Lei de Amdahl.

$$speedup = \frac{1}{(1 - f) + \frac{f}{k \text{ (melhoria)}}}$$

- Usando a Lei de Amdahl, calcule o ganho de velocidade: 67% paralelo com a) 2 núcleos e b) 4 núcleos

# Multicore

- Desafio para o desenvolvedor: identificar partes do código ‘sequencial’ que podem ser ‘paralelizados’
- Threads são de baixo custo computacional, mas baixo nível e gerencia mais cuidadosa – Linux/POSIX (pthreads); Windows (WinThreads)
- APIs (Application Interfaces) como OpenMP e MPI são modelos de alto nível

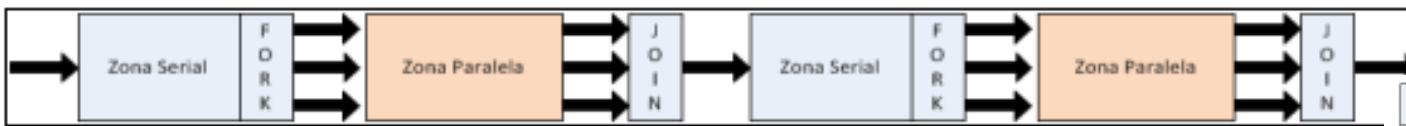


Figura 2.3. Fork-join no OpenMP

```
//soma e imprime a soma do vetor
void somaVetor(double *vetor, int tamanho)
{
    int i =0;
    double soma=0;
    #pragma omp parallel for reduction(+:soma)
    for(i=0; i<tamanho; i++)
    {
        soma+=vetor[i];
    }
    printf("Resultado da soma %3.f\n", soma);
}
```

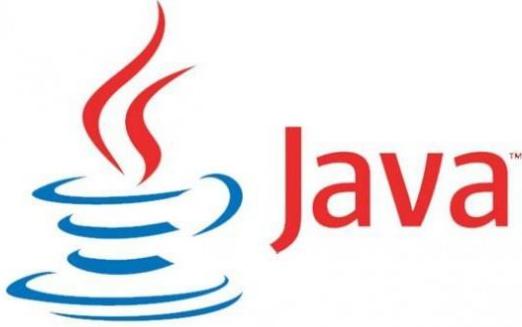
```
#include <stdio.h>
#include <mpi.h>

int main(int argc, char *argv[])
{
    MPI_Init(&argc, &argv);
    // Instruções MPI e a implementação do seu código
    MPI_Finalize();
    return 0;
}
```

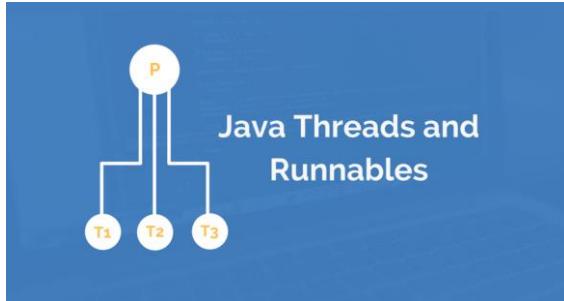
Região paralela do código

Figura 2.6. Estrutura de um código em C/MPI.

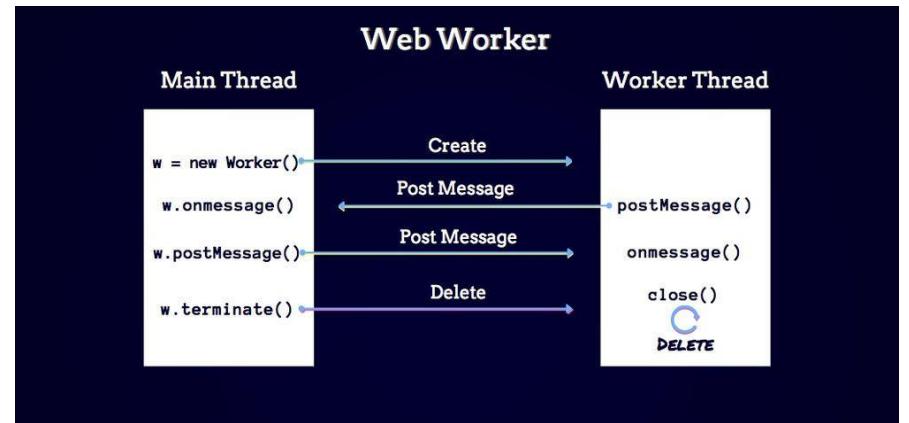
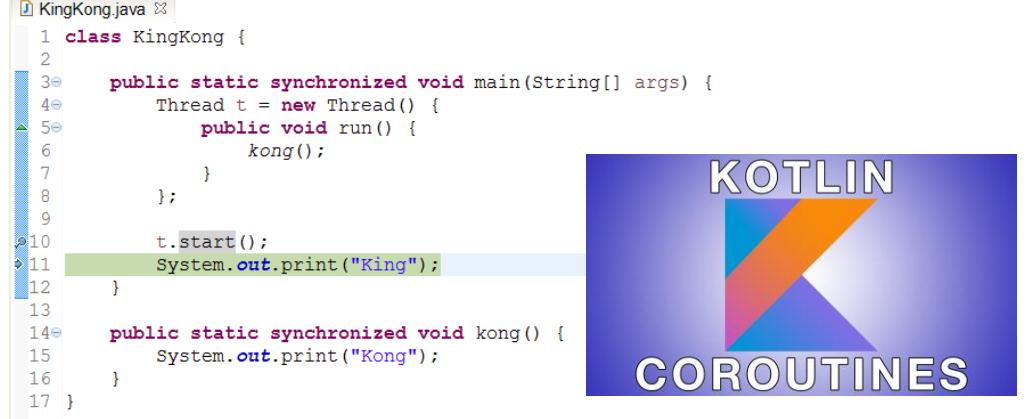
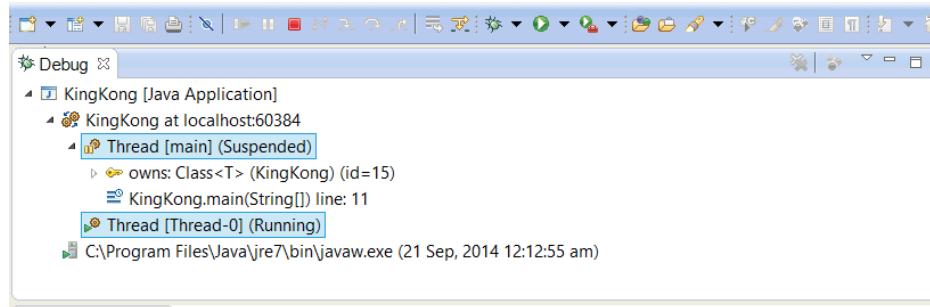
Figura 2.4. OpenMP: a diretiva `#pragma omp parallel for` informa ao compilador que a região deve ser paralelizada.



Observação: não apenas a linguagem Java facilita aplicações *multithread*, mas a Java Virtual Machine é um processo *multithread* que provê agendamento de memória para aplicações Java. Aplicações Java podem, então, se beneficiar diretamente dos recursos *multicore*.



### Stream API:



Javascript

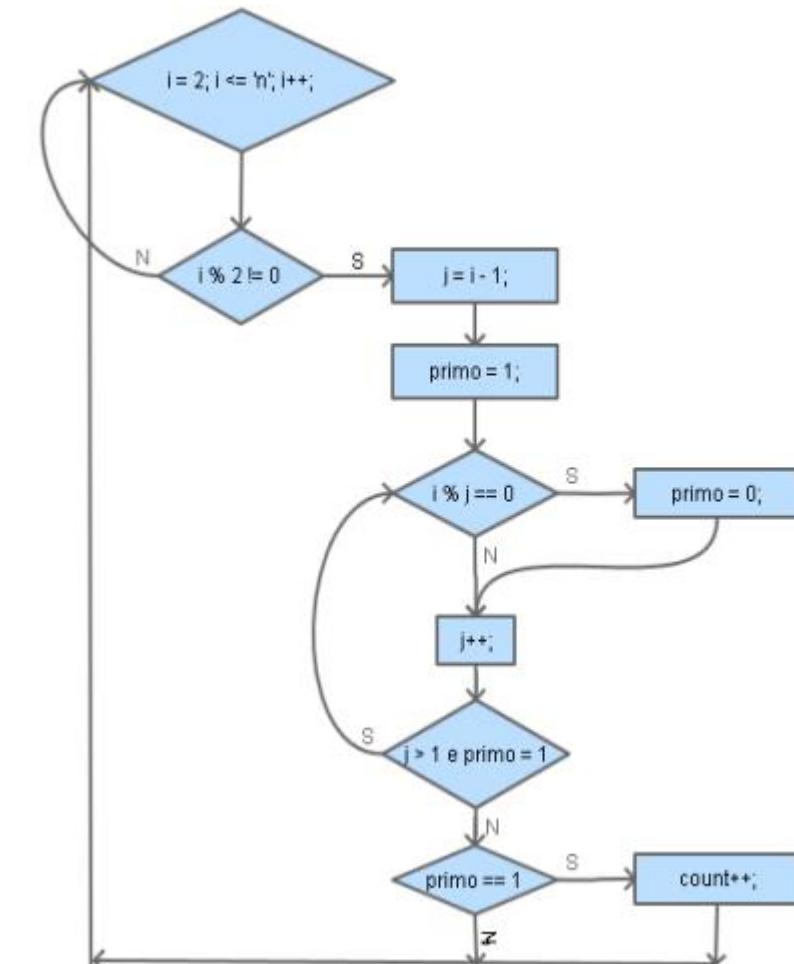
Silva, Cezar G. A.; Asenjo, Maurício N. *Programação paralela – uma introdução ao paralelismo com OPEN MPI*. In: 13. Congresso Nacional de Iniciação Científica, 2013. Disponível em: <<http://conic-semesp.org.br/anais/files/2013/trabalho-1000015028.pdf>>

Com as aplicações rodando em um computador equipado com o processador Intel Core i3 M350x4 em ambiente Linux, foram obtidos os seguintes resultados:

Quadro 2 – Representação dos resultados obtidos em um processador Intel Core i3.

Quantidade de Números	Sem Paralelismo	Paralelo - 2 Processos	Paralelo - 4 Processos
100	0,000028s	1,175111s	1,039873s
1.000	0,002574s	1,027346s	1,047605s
10.000	0,110456s	1,106495s	1,111911s
100.000	10,531448s	8,884402s	7,512204s
1.000.000	1050,243978s	786,315063s	579,89856s

A presença do paralelismo é evidente nos maiores números e também quando levamos o paralelismo aos limites do processador. Durante a execução da aplicação sem paralelismo foi observado que o uso de CPU não ultrapassou os 30%, porém no algoritmo paralelo dividido em quatro processos o uso de CPU atingiu a marca de 100%, ou seja, o processador estava trabalhando com carga máxima durante o paralelismo.



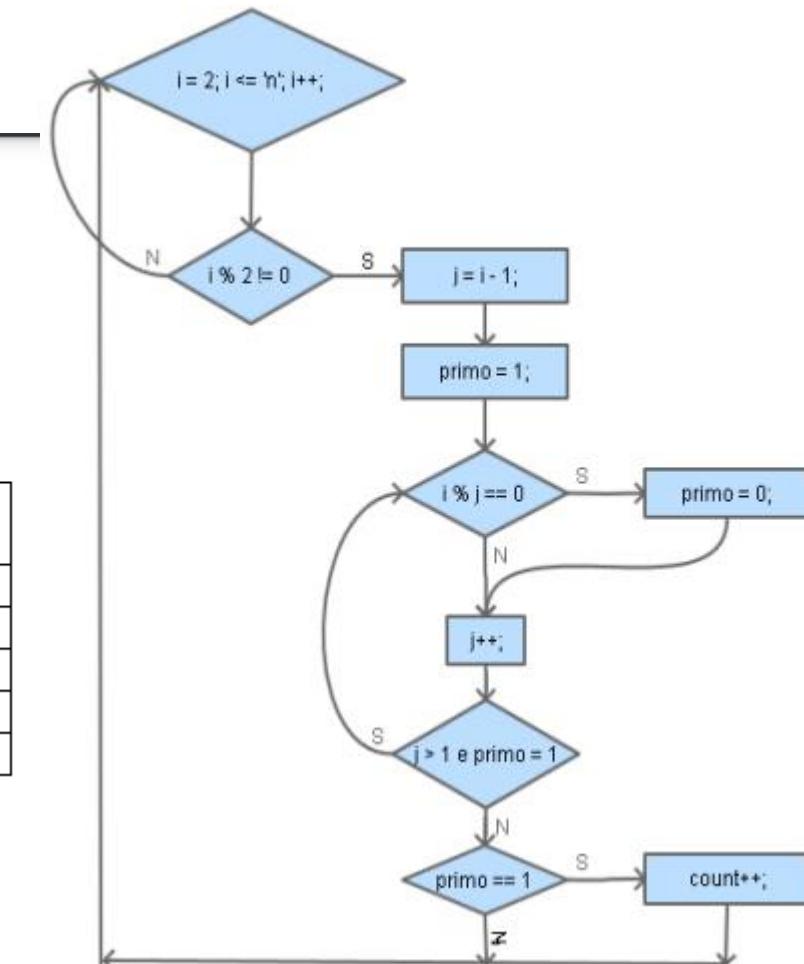
Encontrar o número de primos numa determinada faixa

Silva, Cesar G. A.; Asenjo, Maurício N. *Programação paralela – uma introdução ao paralelismo com OPEN MPI*. In: 13. Congresso Nacional de Iniciação Científica, 2013. Disponível em: <<http://conic-semesp.org.br/anais/files/2013/trabalho-1000015028.pdf>>

Para ter uma melhor comparação o teste foi realizado também sobre outra configuração, um computador equipado com um processador Intel Core i7x8 em um ambiente Linux rodando em uma máquina virtual. E os resultados obtidos foram:

**Quadro 3 – Representação dos resultados obtidos em um processador Intel Core i7.**

Quantidade de Números	Sem Paralelismo	Paralelo - 4 Processos	Paralelo - 8 Processos
100	0,000012s	1,060125s	1,142675s
1.000	0,000467s	1,065693s	1,089045s
10.000	0,106573s	1,083212s	1,867748s
100.000	9,743917s	5,705782s	4,066512s
1.000.000	959,056274s	496,121704s	373,259491s



Encontrar o número de primos numa determinada faixa