Processo de fabricação VLSI e breve introdução a arrays sistólicos TPU

MAC0344 - Arquitetura de Computadores
Prof. Siang Wun Song

Slides usados: https://www.ime.usp.br/∼song/mac412/vlsi-fab.pdf

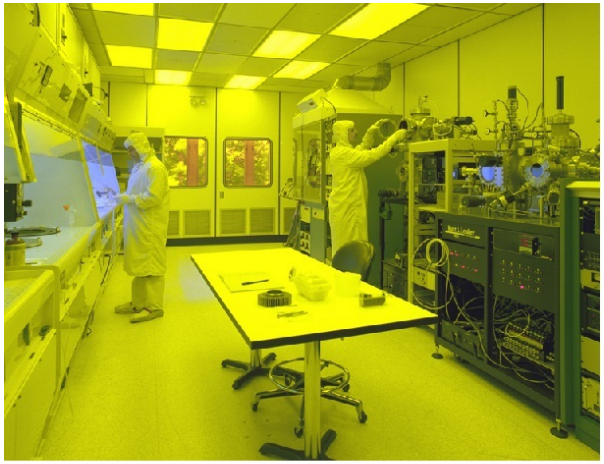
Baseado em parte em Mead and Conway - Introduction to VLSI Systems
Esse assunto não cai em provas

Fabricação de chips VLSI e Arrays Sistólicos

- Fabricação de chips VLSI e conceito de arrays sistólicos
- Ao final desta aula vocês saberão
  - O processo básico para fabricação VLSI
  - Pastilhas VLSI podem ser fabricadas para aplicações específicas (ASICs), por exemplo usando FPGAs (Field Programmable Gate Arrays) ou arrays sistólicos.
  - Um exemplo de um array sistólico para multiplicar duas matrizes.
  - Google TPU é um array sistólico usado em Google Search, Google Street View, Google translate para computações de redes neurais artificiais em aprendizado de máquina.

Esse assunto não cai em provas.

Fabricação de pastilhas VLSI



Source: Wikipedia

- Instalações de alto custo (TSMC Taiwan investiu 9 bilhões de dólares e planeja uma fábrica de 20 bilhões)
- Ambiente urbano: 35 milhões de partículas de 0,5 μm por m^3.
- Sala limpa ISO 1: $\leq$ 12 partículas de 0,3 μm por m^3. Mais exigente do que uma sala cirúrgica.
- Controle de temperatura e humidade.
- Controle contra vibração - equipamentos ou uma sala inteira colocada em cima de isolador de vibração.

Fonte: Wikipedia - Semiconductor fabrication plant.

Processo básico

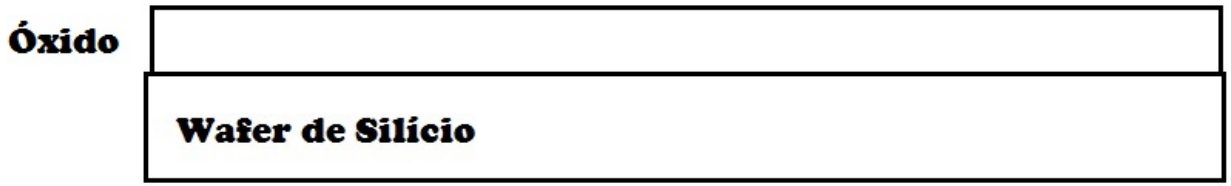
Source: PublicDomainPictures.net

- Veremos o processo básico.
- O processo básico usa uma máscara que possui partes transparentes e partes opacas.
- O processo básico produz a forma da máscara na pastilha de silício.
- A máscara é parecida com o negativo de fotografia e serve para levar a imagem do negativo para um papel fotográfico.

Processo Básico

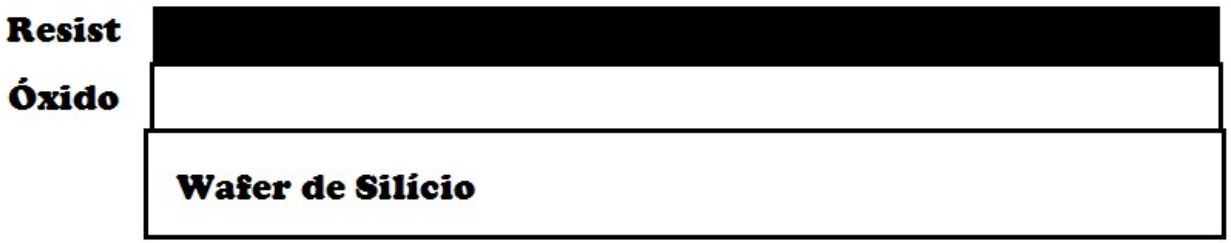
- Expor wafer de silício a oxigênio num forno de alta temperatura.

Processo Básico



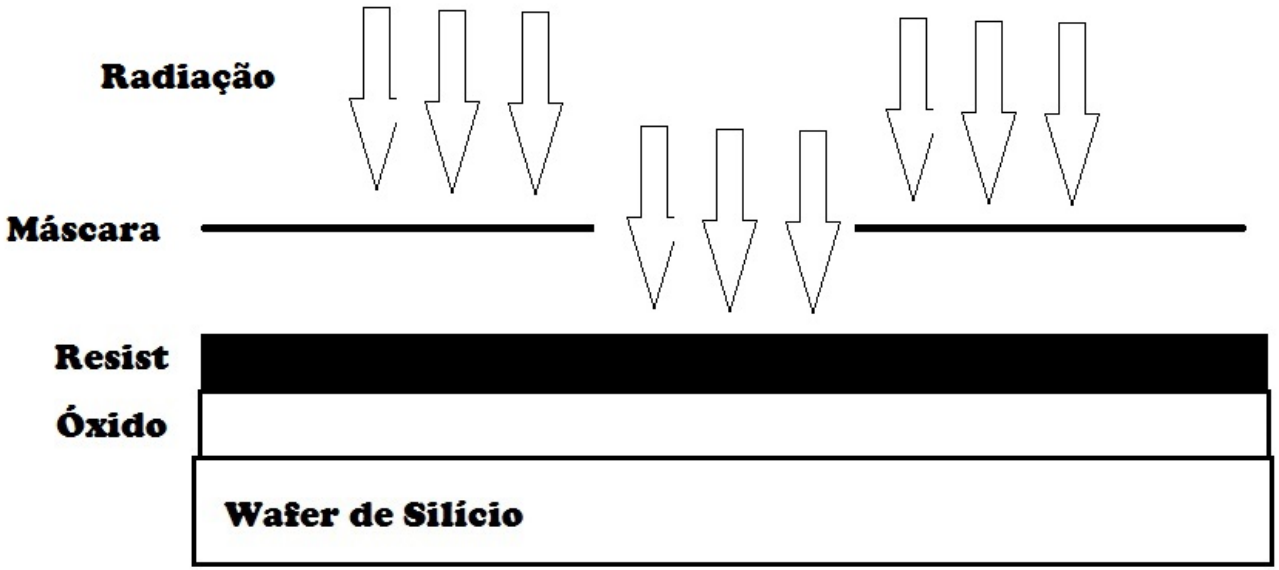
- Expor wafer de silício a oxigênio num forno de alta temperatura. Forma-se óxido S_iO_2 na superfície.

Processo Básico



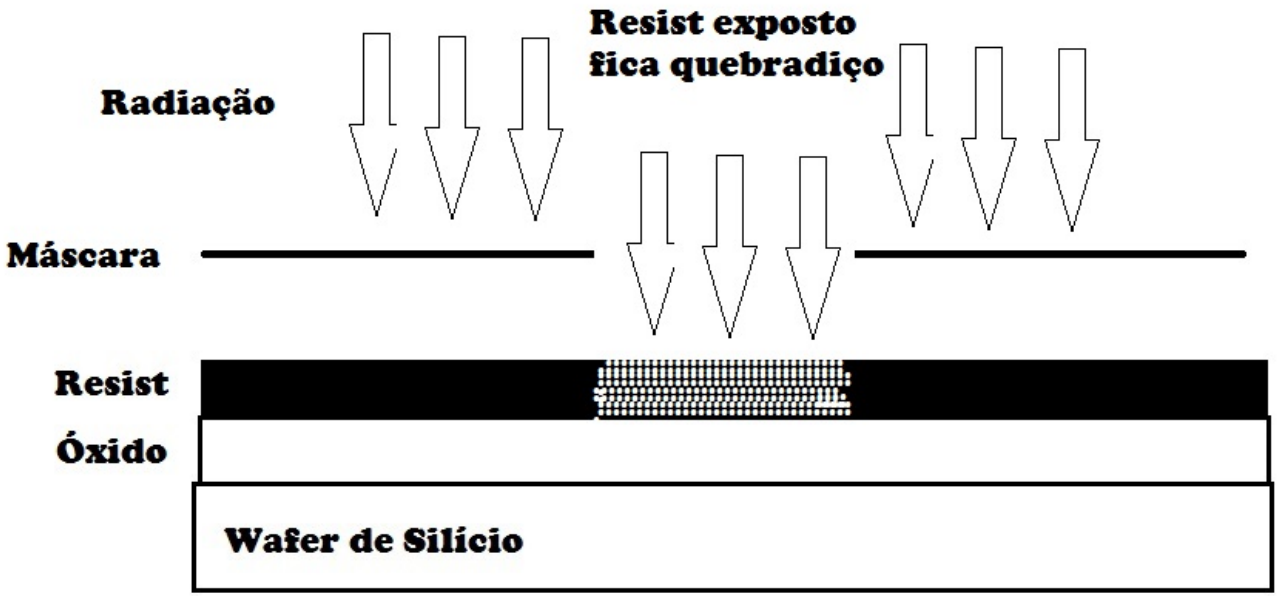
- Pintar com uma fina camada de material orgânico chamado “resist”. Secar e “assar” no forno.

Processo Básico



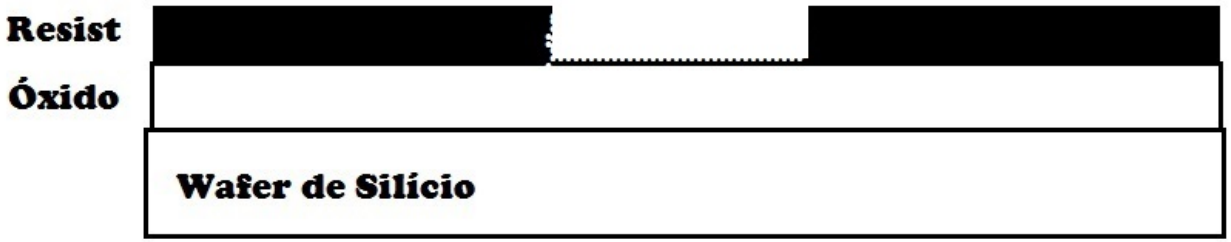
- Incidir radiação intensa de luz ultravioleta ou raio-X através de uma máscara.

Processo Básico



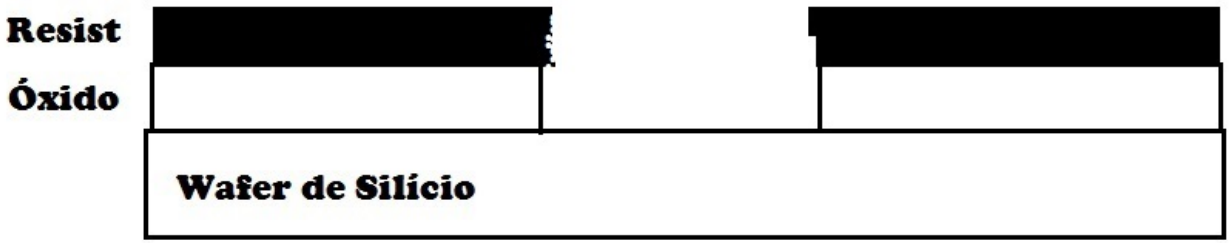
- Isso vai quebrar a estrutura de moléculas de parte (expostas) do resist.

Processo Básico



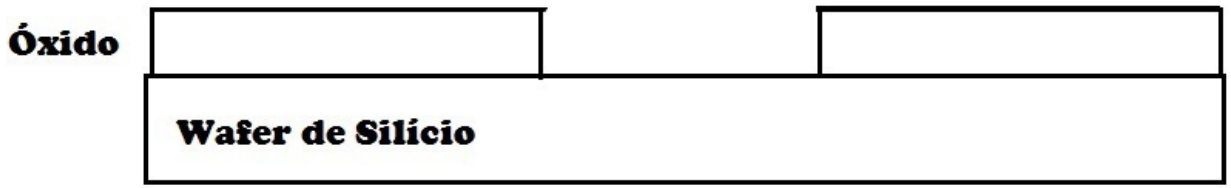
- Usar banho de solvente para tirar “resist”expostos (quebradiços).

Processo Básico



- Usar ácido hidrofluórico que dissolve o óxido S_iO_2 mas não ataca o “resist”.

Processo Básico



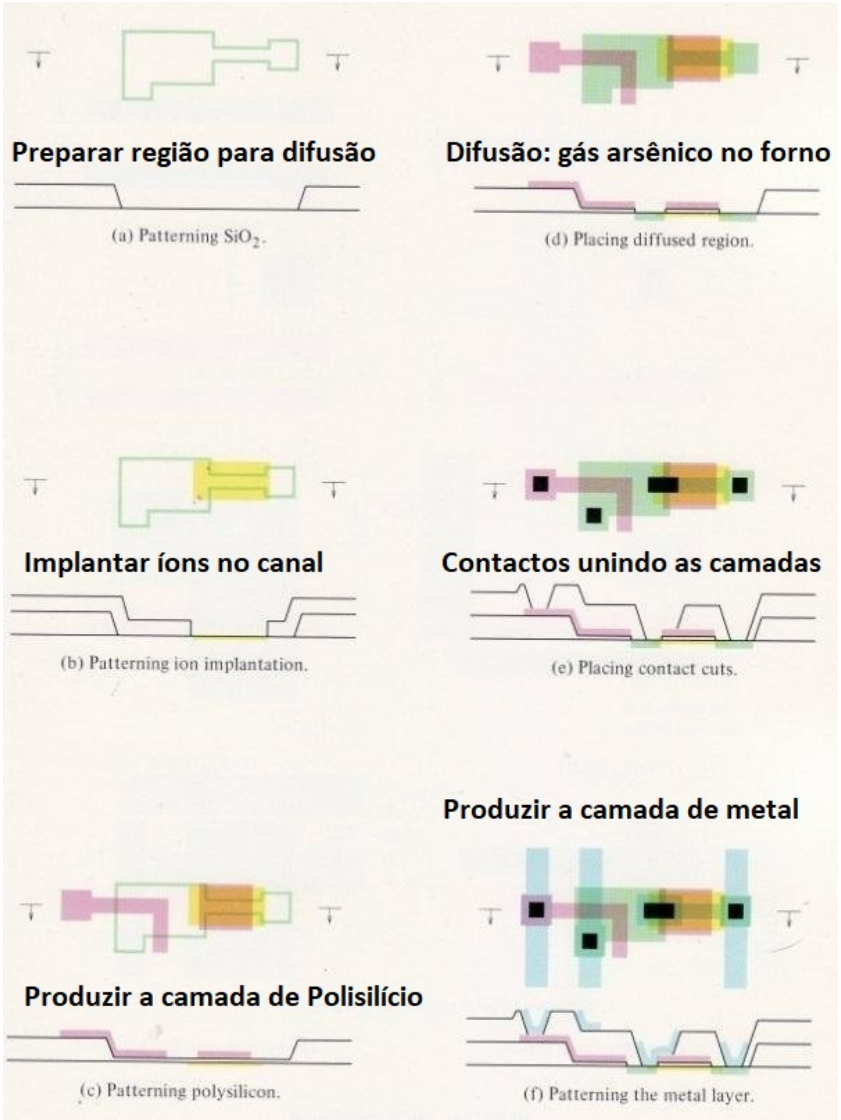
- Eliminar “resist”com solventes fortes ou ácidos. O processo básico produz a forma da máscara no chip.

Processo Completo

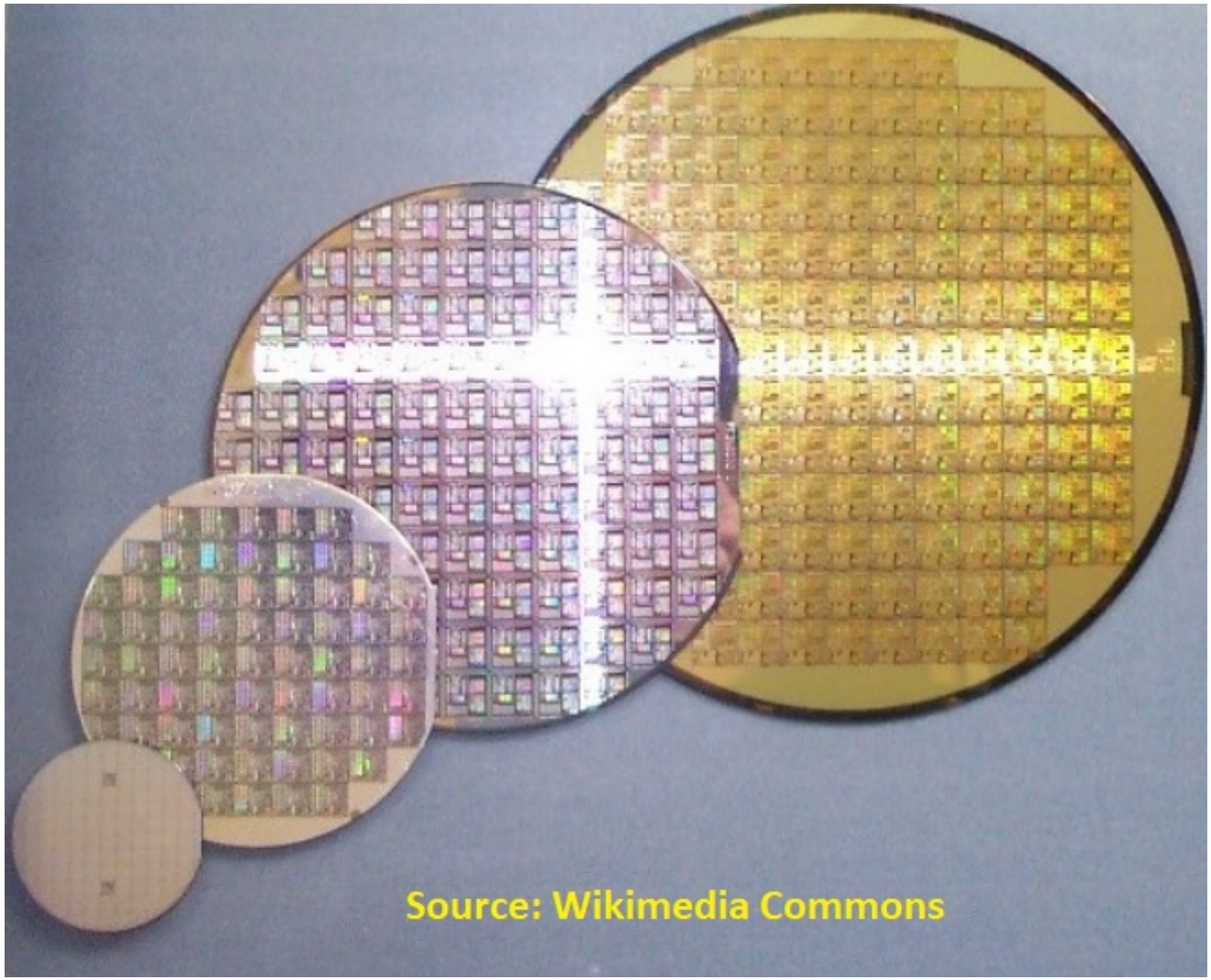
- O processo básico produz a forma da máscara no chip.
- Lentes de redução colocadas entre a máscara e o wafer reduzem a forma da máscara ao chegar no wafer.
- O processo básico é usado no processo completo para produzir as várias camadas (difusão, poli-silício, metal, etc.) na pastilha, conforme as respectivas máscaras.
- O próximo slide mostra as etapas do processo completo NMOS.

Processo Completo

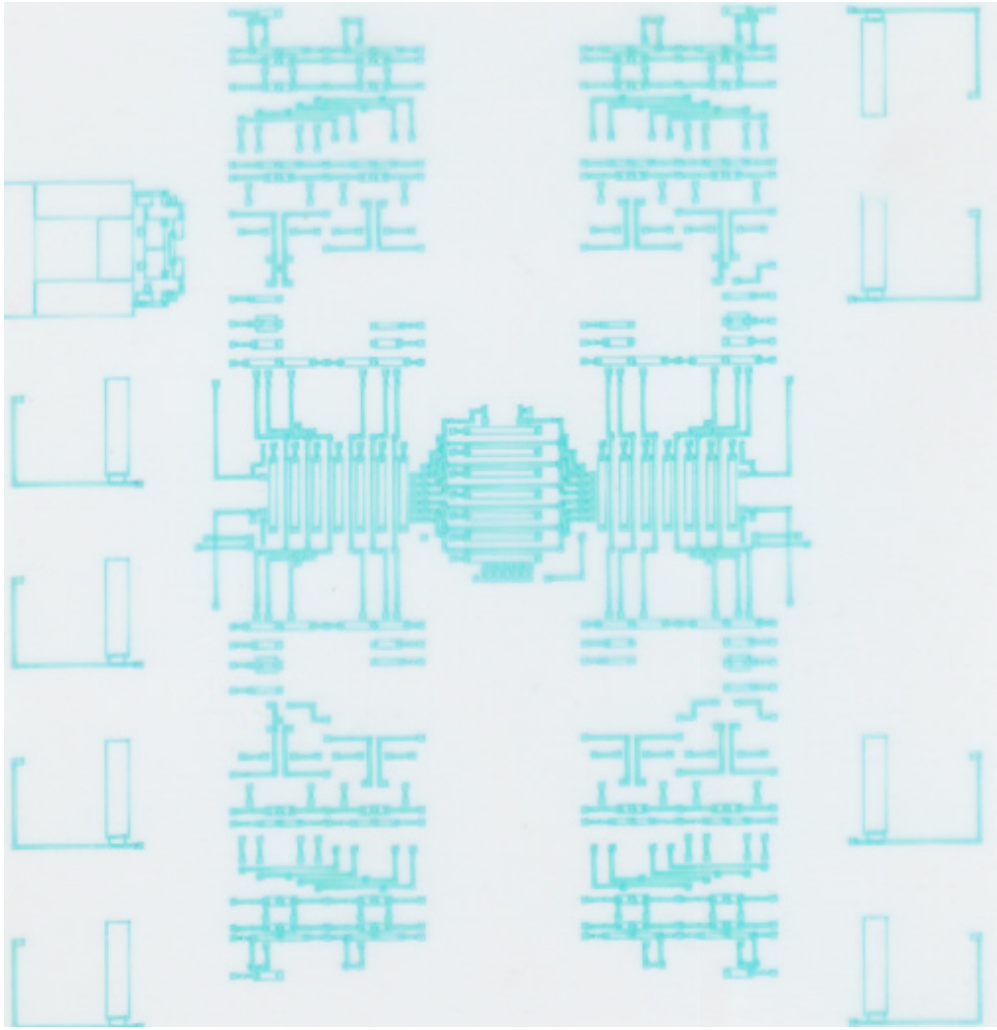
Source: Mead and Conway - Introduction to VLSI Systems



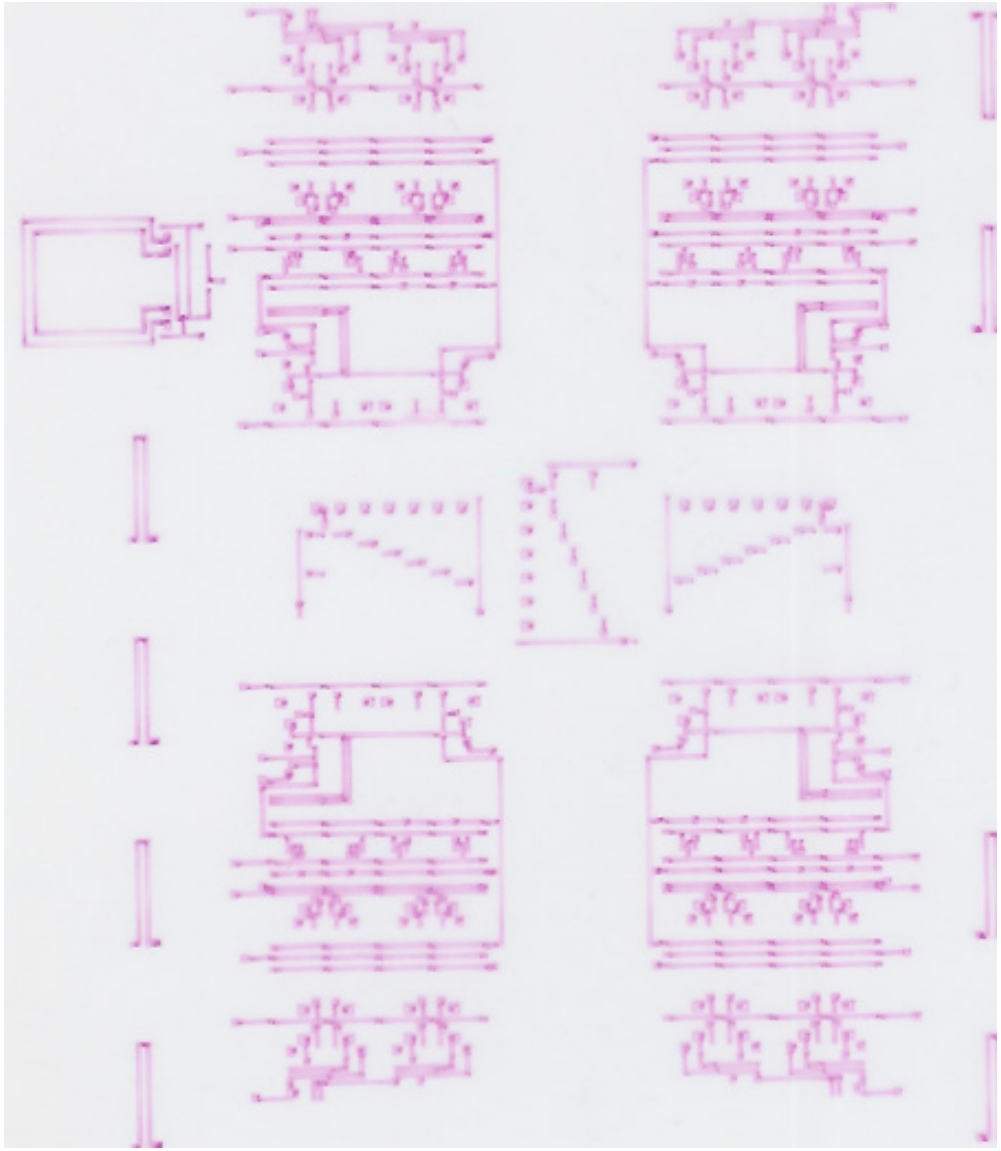
Wafers de Silício



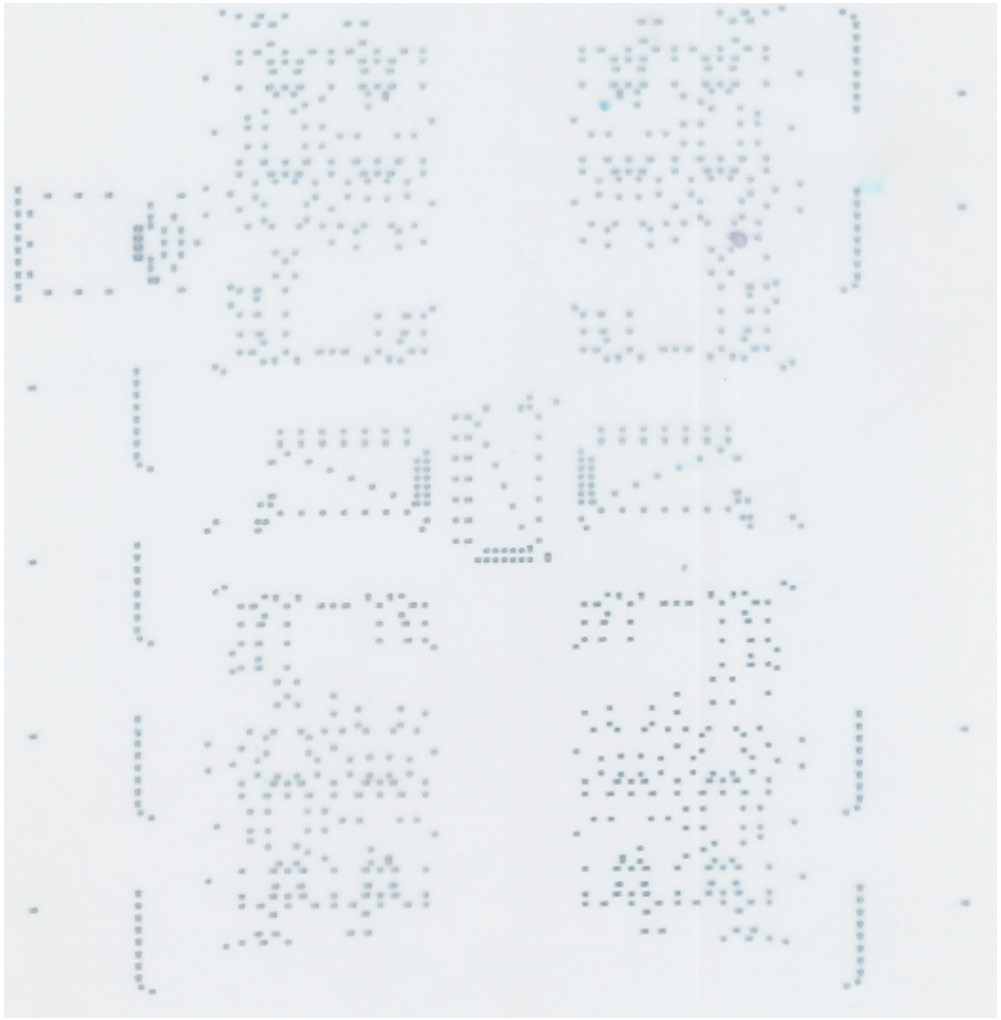
Máscara Difusão



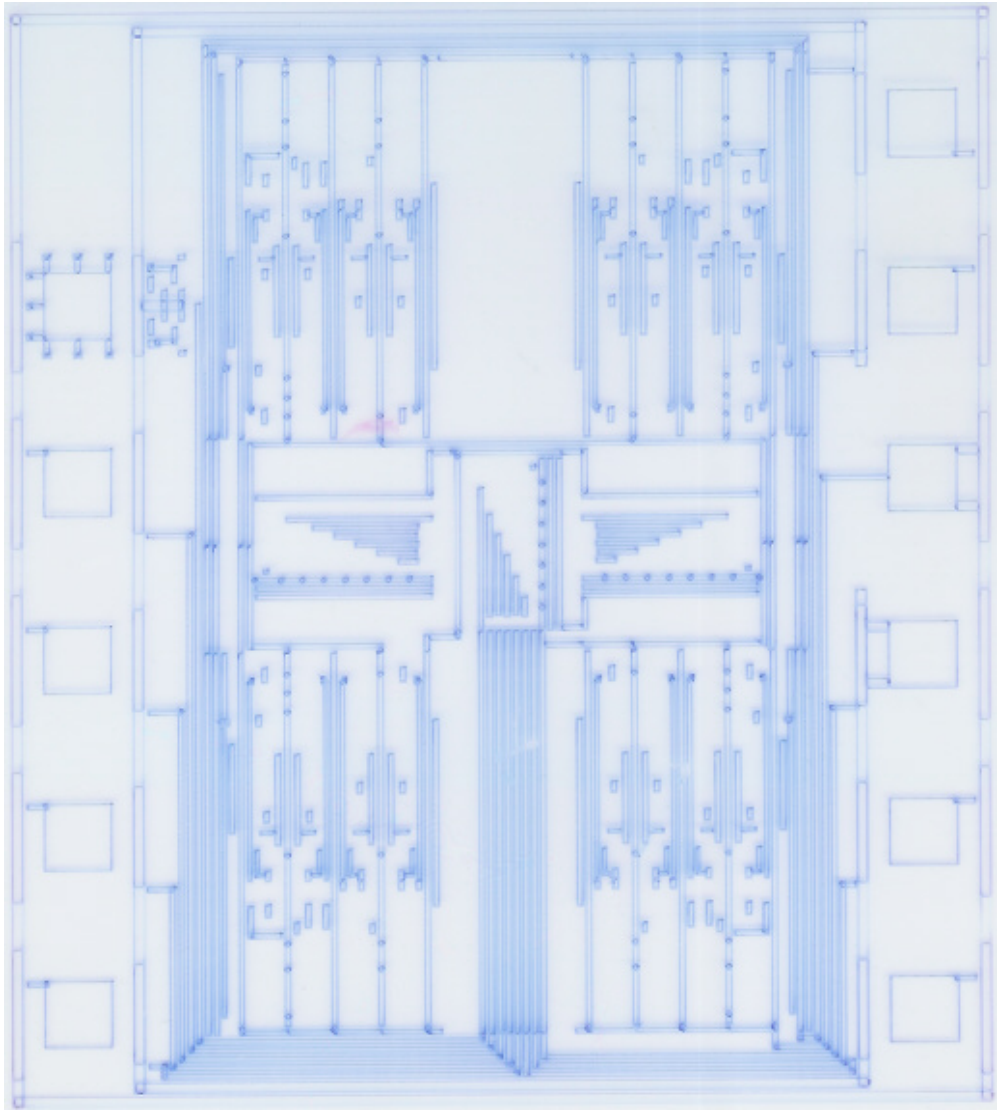
Máscara Poli-silício



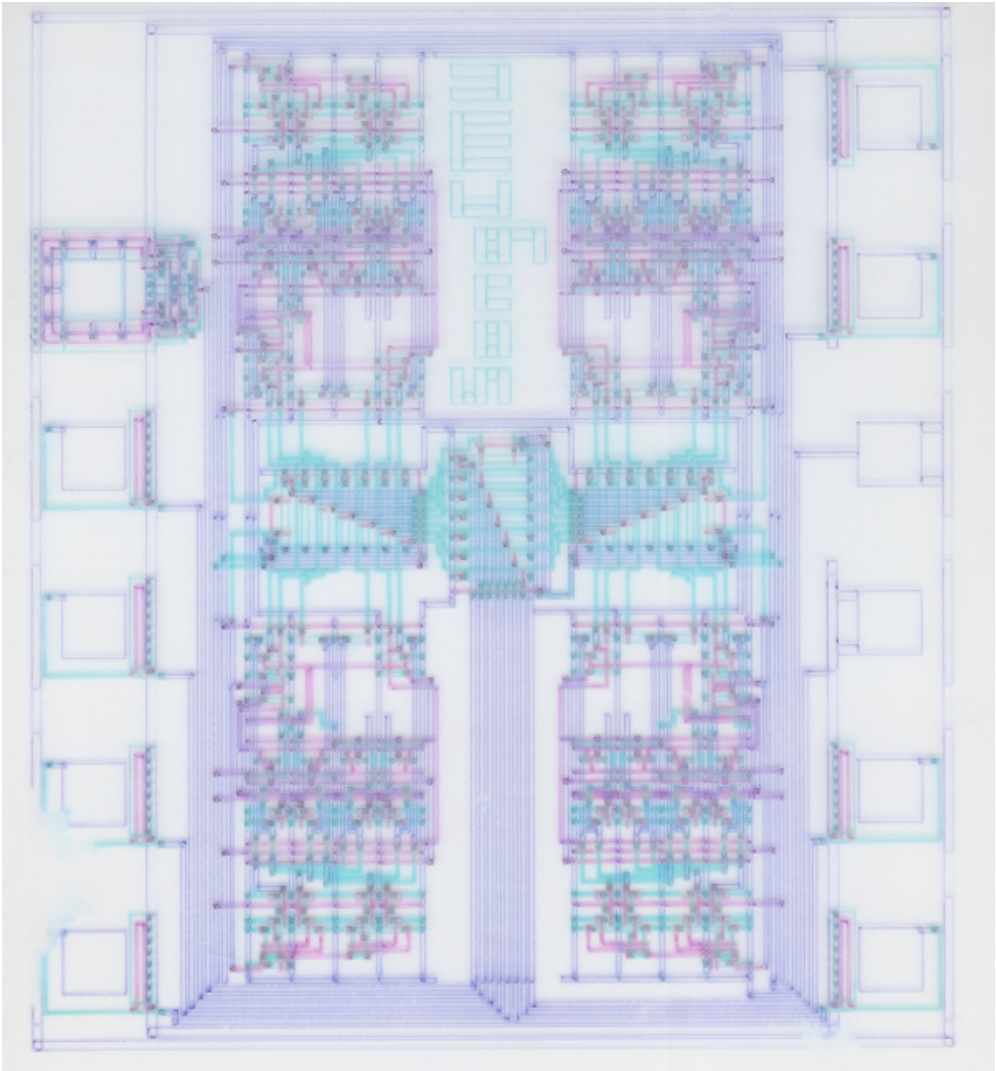
Máscara Contatos



Máscara Metal



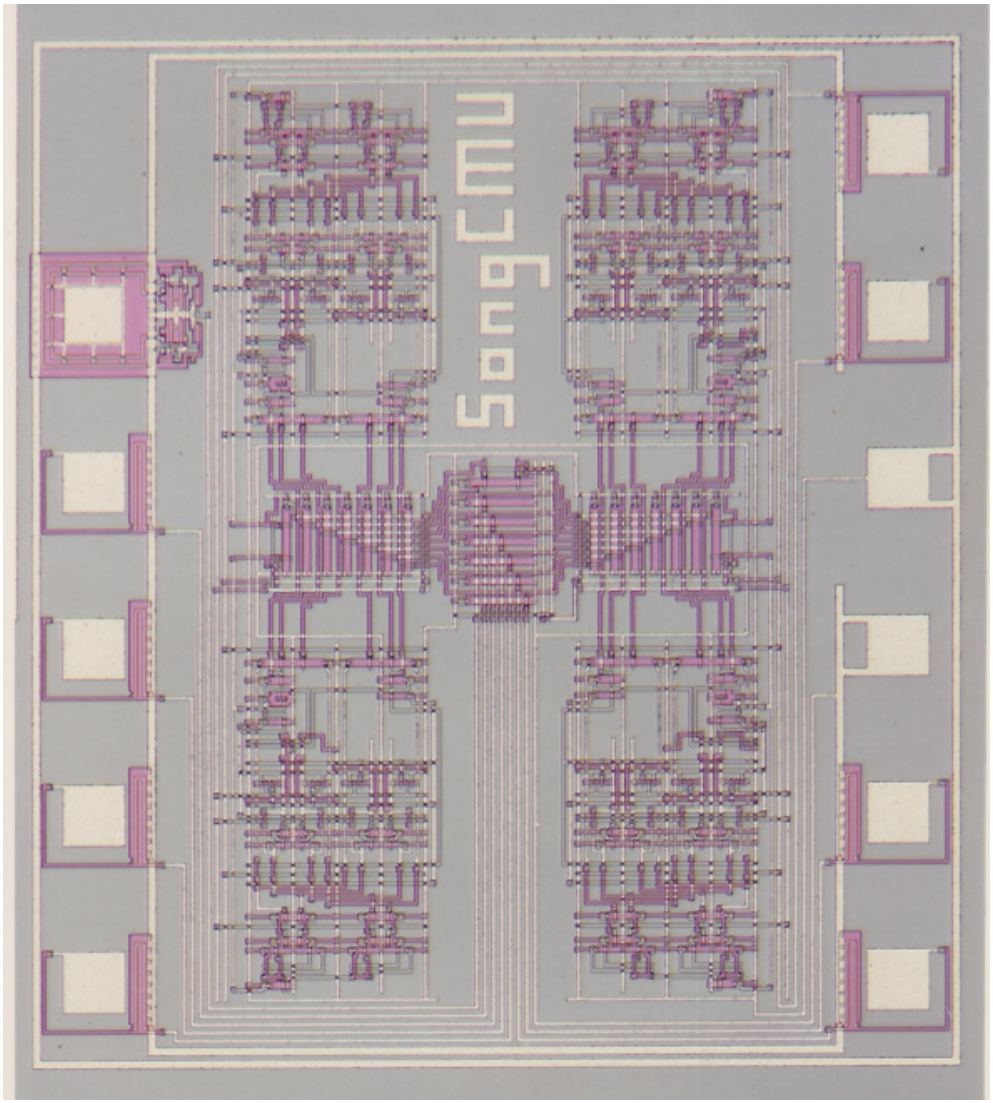
Todas As Camadas Juntas



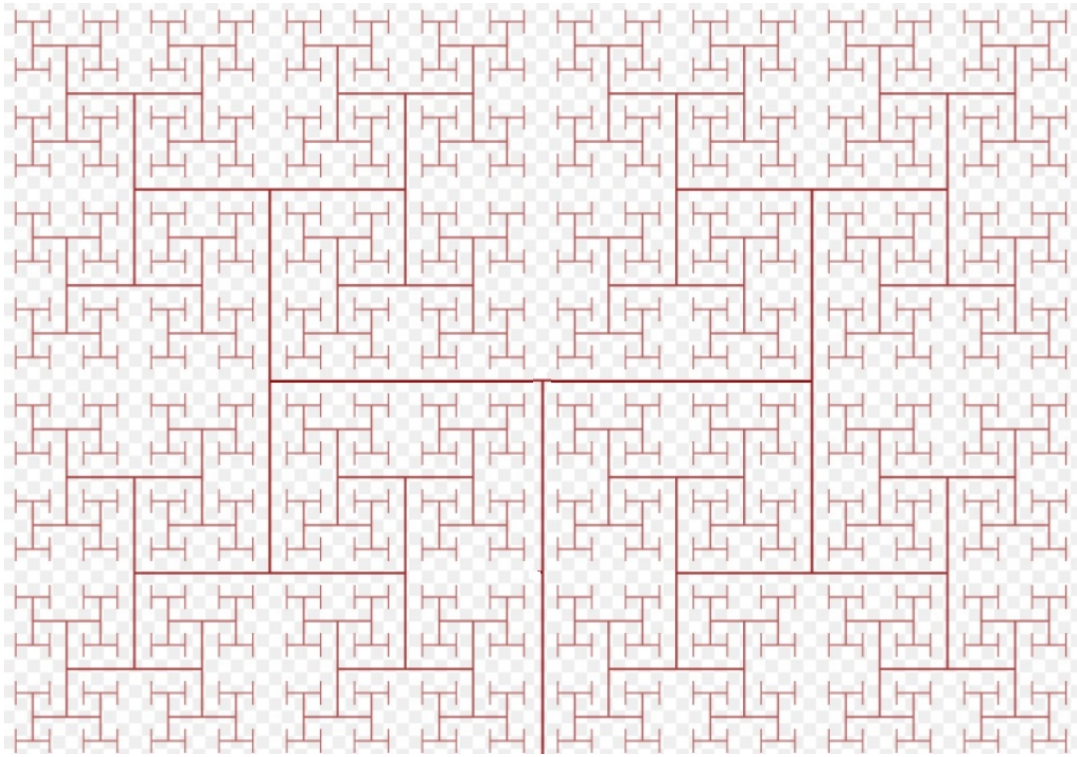
Projeto de pastilha customizada para aplicação específica

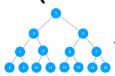
- A tecnologia VLSI é usada para processadores e memória.
- Propicia também o projeto de pastilhas customizadas para aplicações específicas ou ASICs (Application Specific Integrated Circuits). Anualmente há a conferência ASAP (International Conference on Application-specific Systems, Architectures and Processors).
- Tipicamente isso é feito usando FPGAs (Field Programmable Gate Arrays).
- ASICs podem também ser projetados com o método de Arranjos Sistólicos (Systolic Arrays) propostos nos anos 80.
- O Systolic Array consiste de um conjuntos de células (elementos de processamento) simples interconectadas de uma forma regular no plano.

Projeto busca por árvore binária

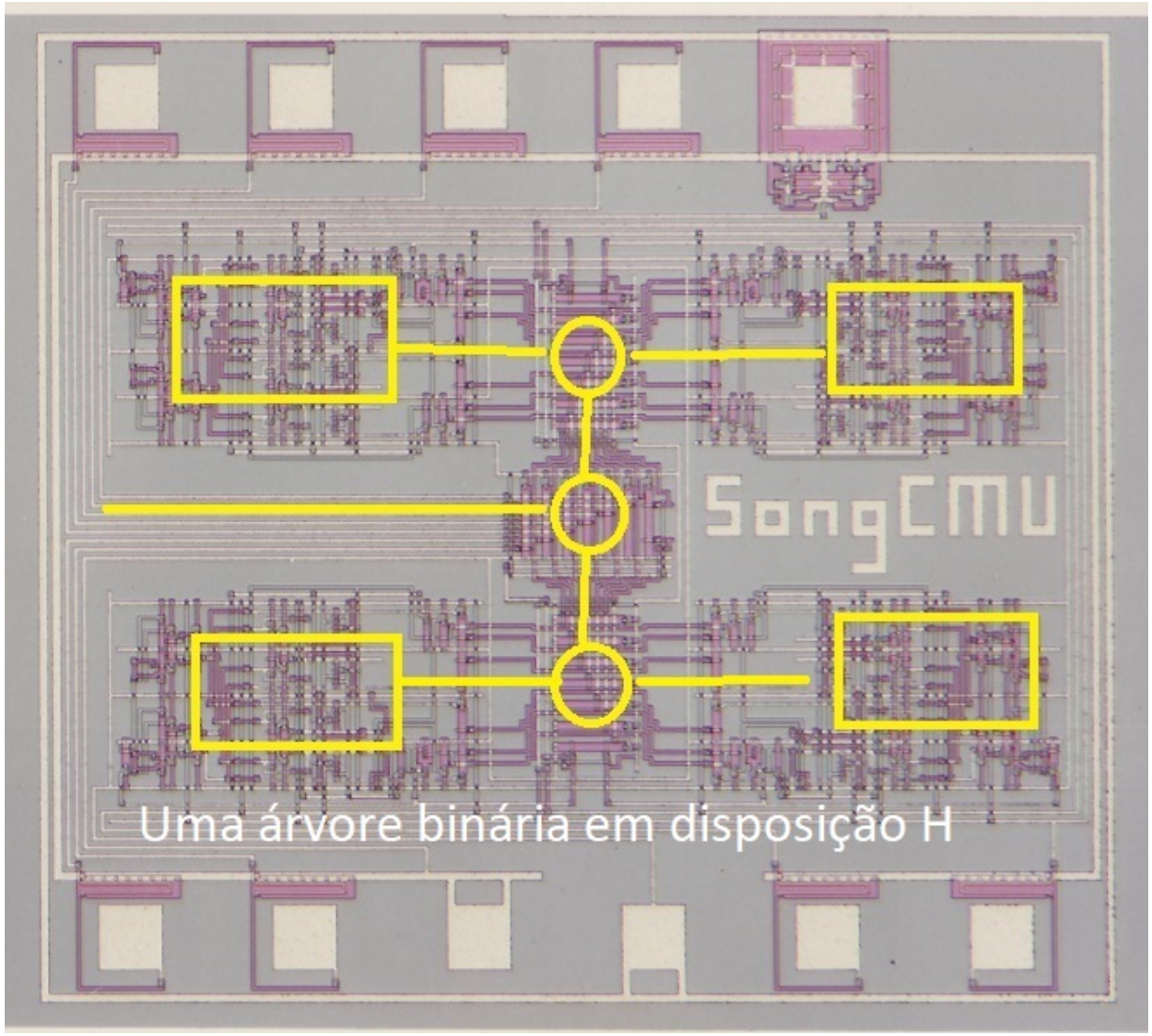


Disposição-H de uma árvore binária no plano

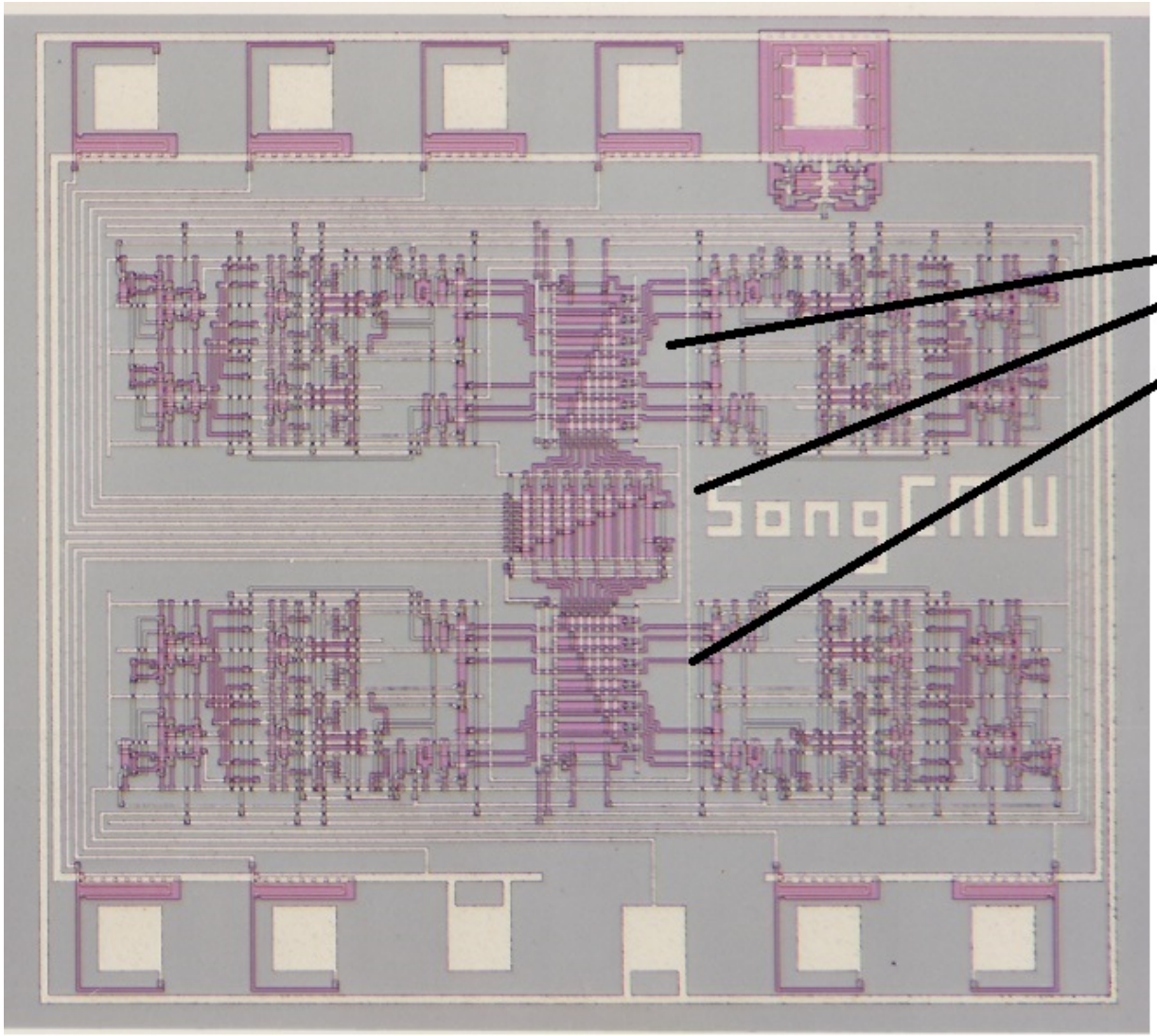


- Acima mostramos uma árvore binária em disposição-H (nome devido à forma H que aparece no desenho) que melhor utiliza o espaço.
- Quantos nós tem essa árvore acima? (Tente desenhá-la na forma usual de representar uma árvore binária (i.e.: ) no mesmo espaço acima :-)

Projeto busca por árvore binária



Projeto busca por árvore binária

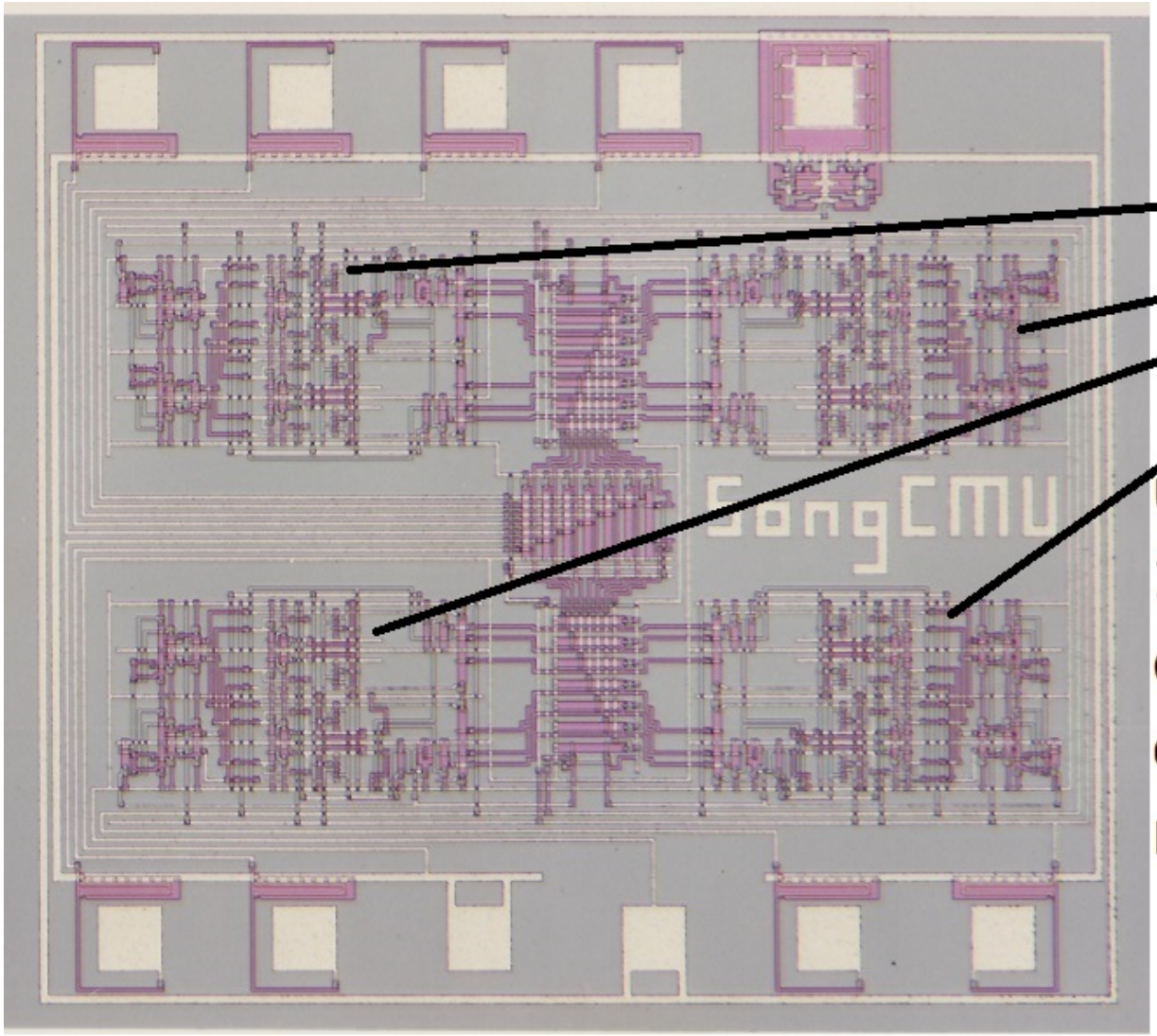


Nós internos: raiz e os seus 2 filhos.

Note que cada filho é igual a raiz mas com uma rotação de 90 graus.

Portanto basta projetar um nó interno.

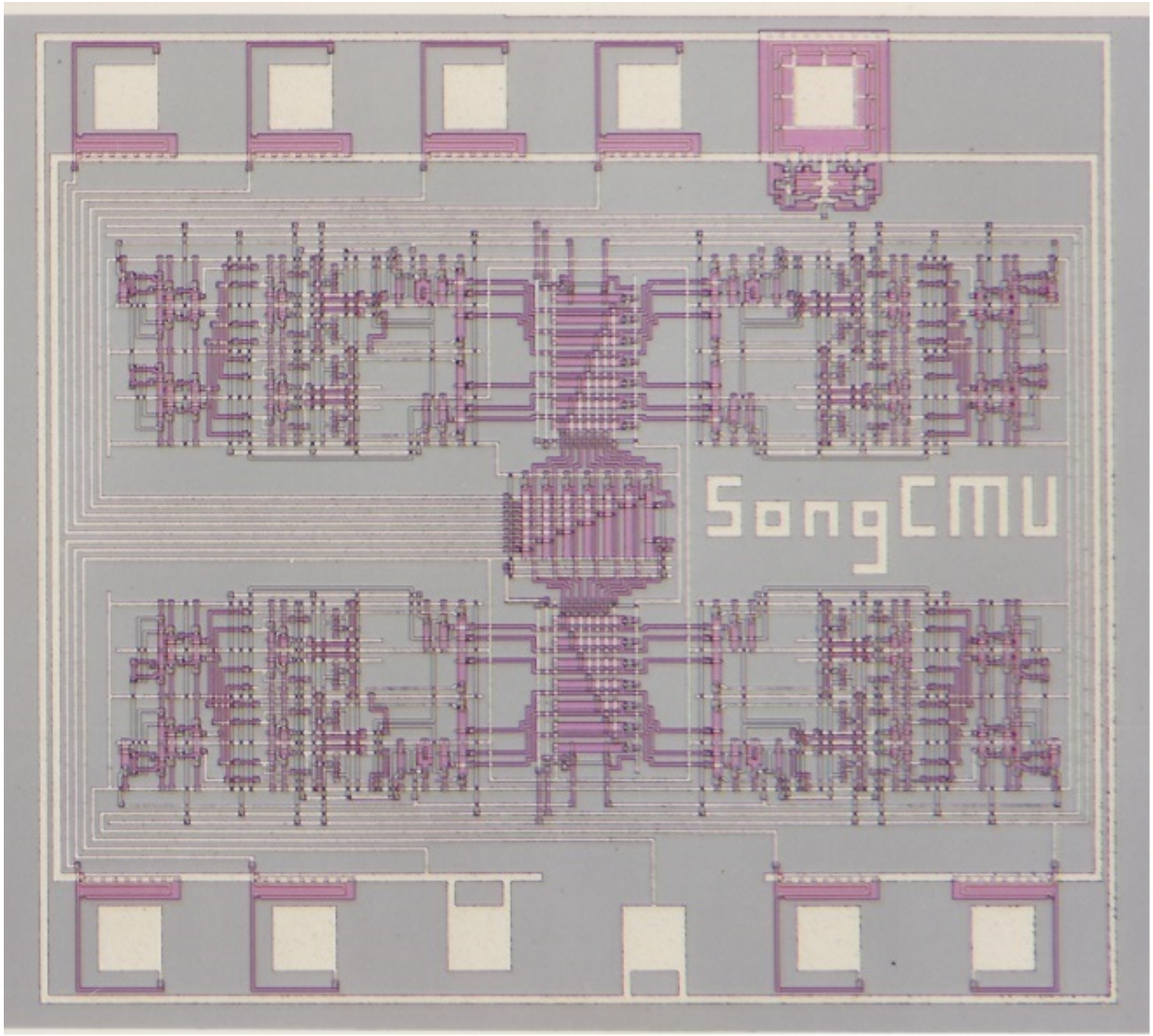
Projeto busca por árvore binária



Nós externos: 4 folhas "idênticas" a menos de rotação.

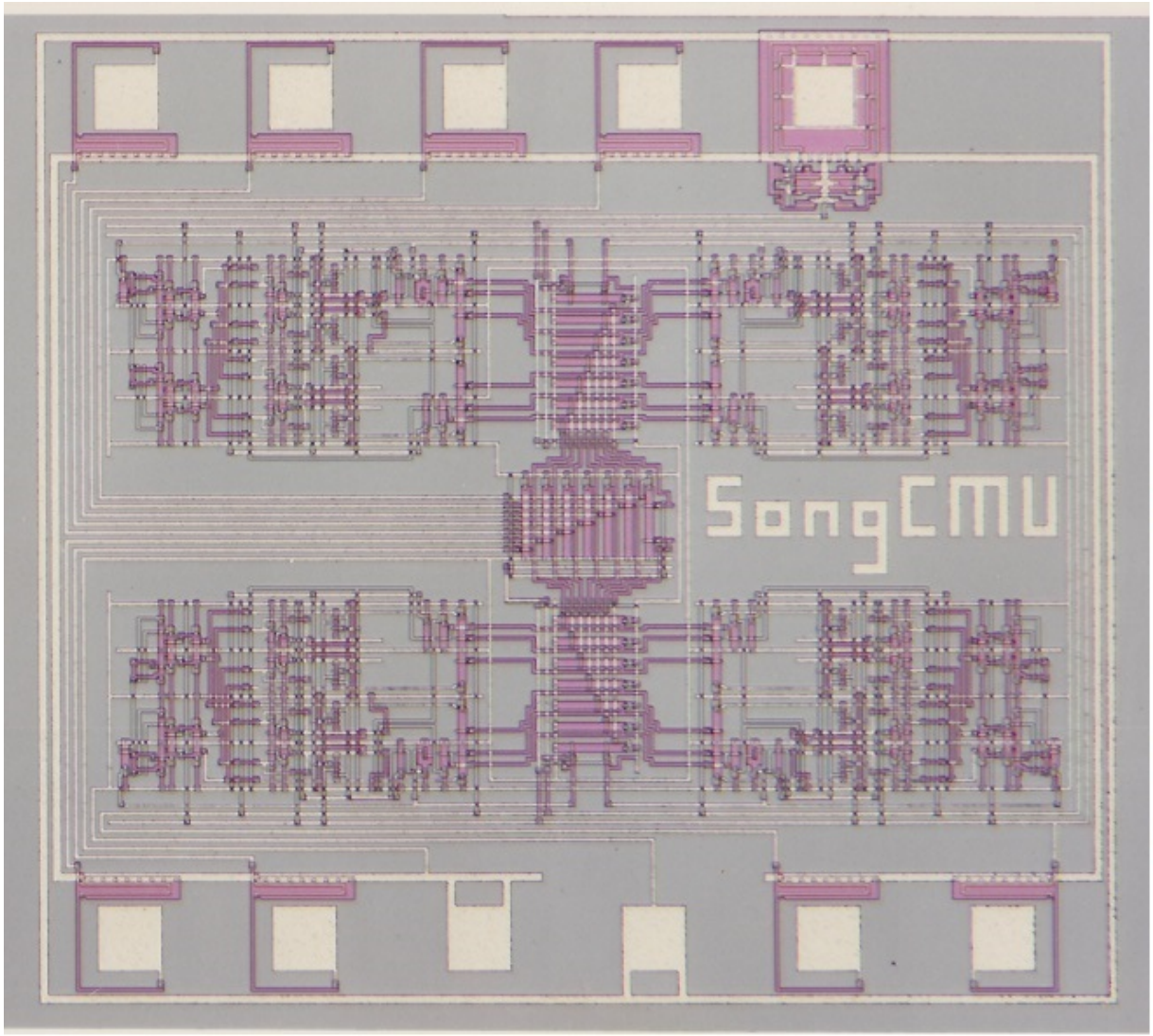
Cada folha tem um endereço 00, 01, 10, 11, uma memória de 2 bits, e um comparador de 2 números de 2 bits.

Projeto busca por árvore binária



Há duas instruções:
. Insere dd em ee
(dd dado de 2 bits e
ee endereço 2 bits)
A instrução é propa-
gada a todas as
folhas. Cada folha
compara ee e se
igual ao endereço
guarda dd na sua
memória.

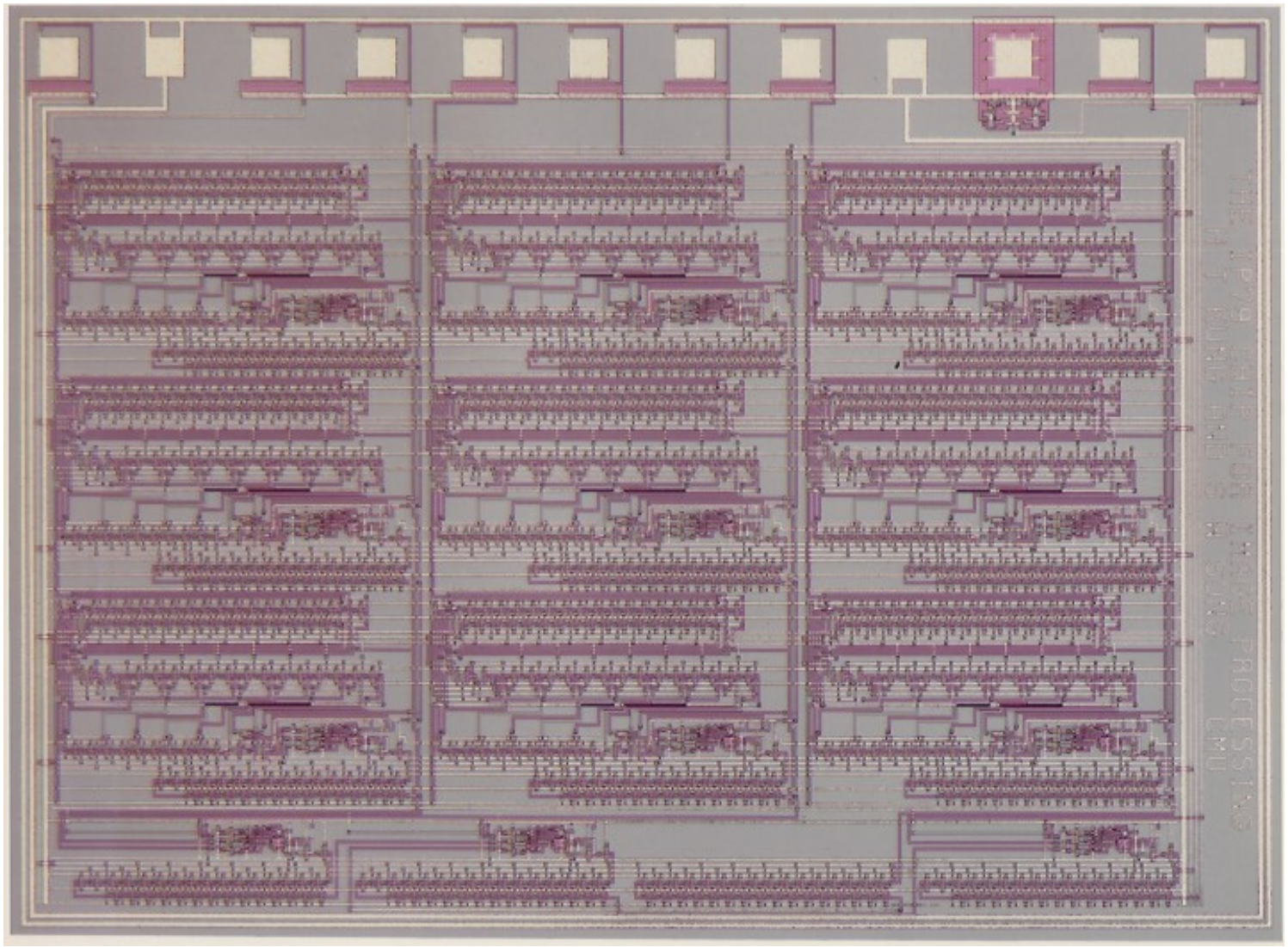
Projeto busca por árvore binária



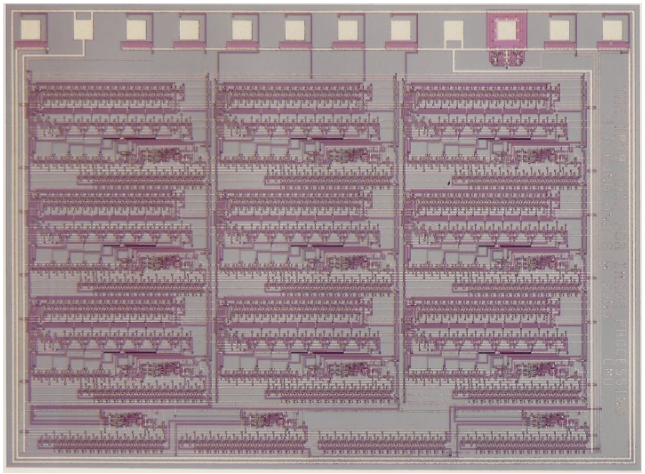
. Busca dd (pergunta se o dd está presente)

A instrução é propagada a todas as folhas. Cada folha verifica se contém dd. Se sim envia 1 senão 0. Cada nó interno faz OR e a raiz devolve resposta

Projeto convolução 2D



Projeto convolução 2D

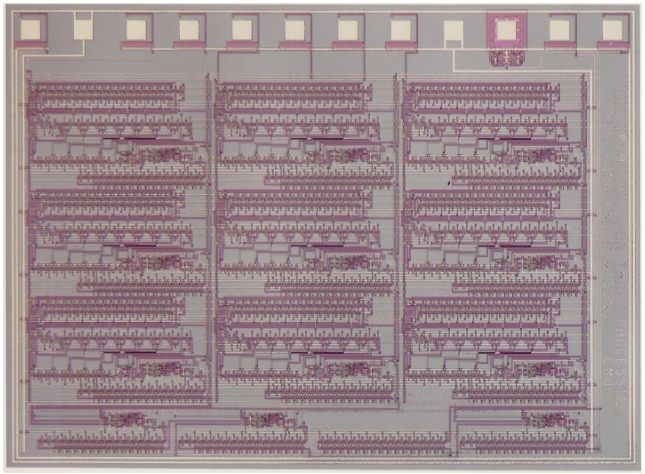


Dada uma matriz 3×3 de pesos: $\begin{pmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{pmatrix}$

Dada uma entrada 3×3 de dados: $\begin{pmatrix} 7 & 6 & 5 \\ 6 & 4 & 3 \\ 5 & 3 & 2 \end{pmatrix}$

A convolução 2d faz a soma ponderada de cada dado multiplicado pelo respectivo peso gerando um número 2.

Projeto convolução 2D

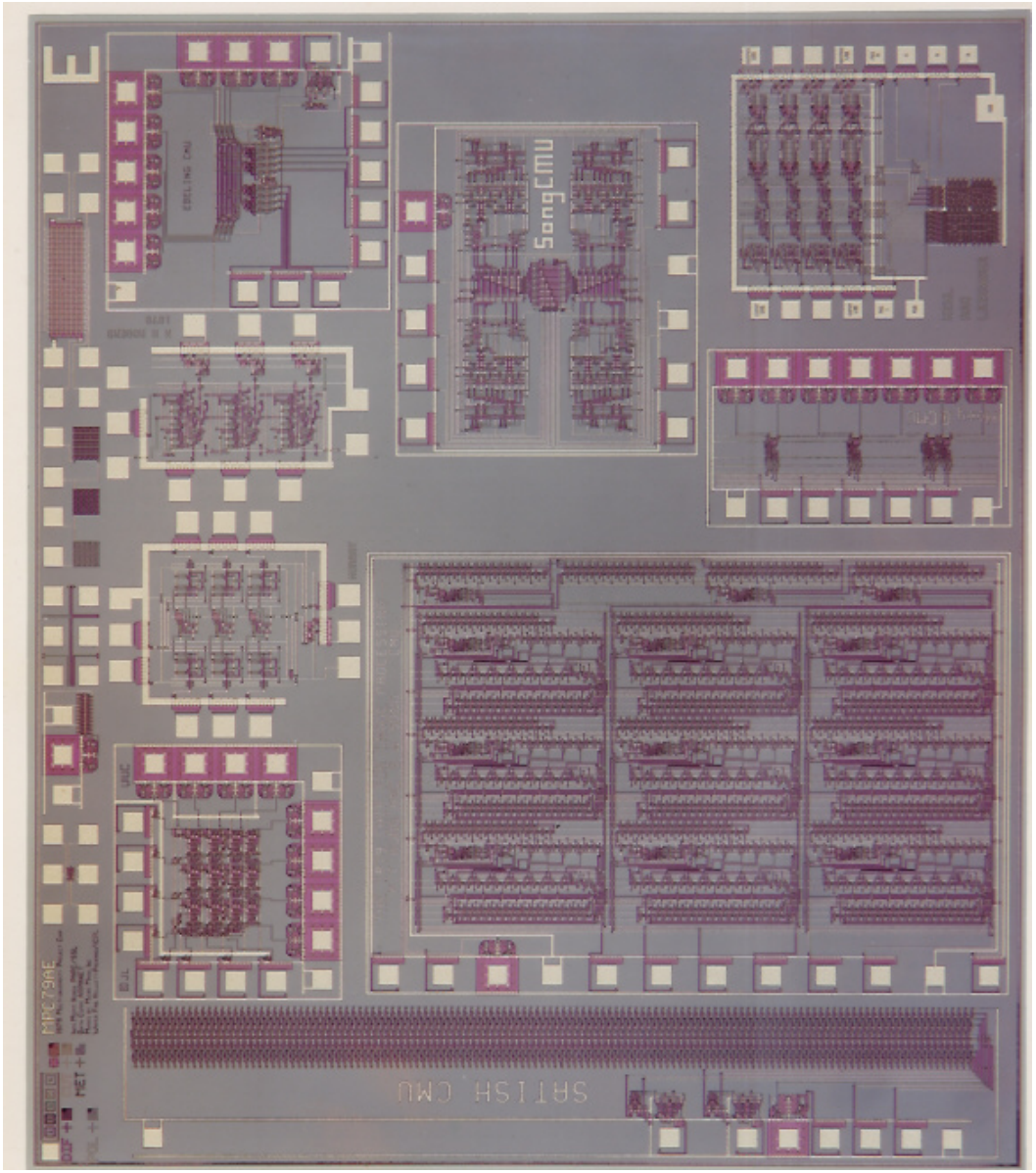


Dada uma matriz K $k \times k$ de pesos e dada uma matriz M $n \times n$, a convolução 2D calcula, para cada posição da matriz K sobre a matriz M, a soma ponderada mencionada acima, gerando uma nova matriz $n \times n$.

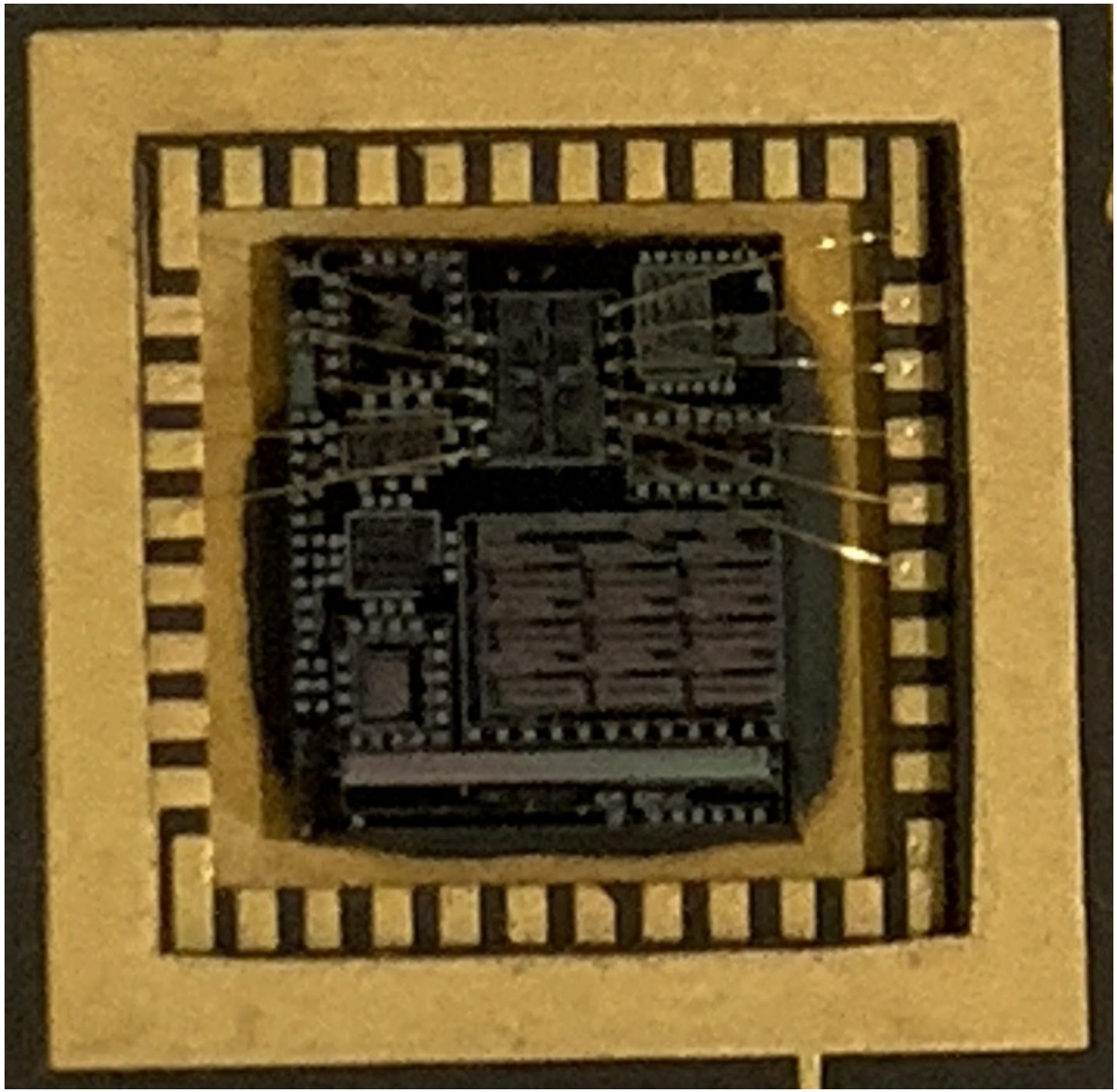
Ver uma animação: 2D Convolution animation gif.

A convolução é usada em Redes Neurais Artificiais para Aprendizado em Máquina e uma implementação eficiente é por meio de multiplicação de matrizes. Ver 2D Convolution as a Matrix-Matrix Multiplication.

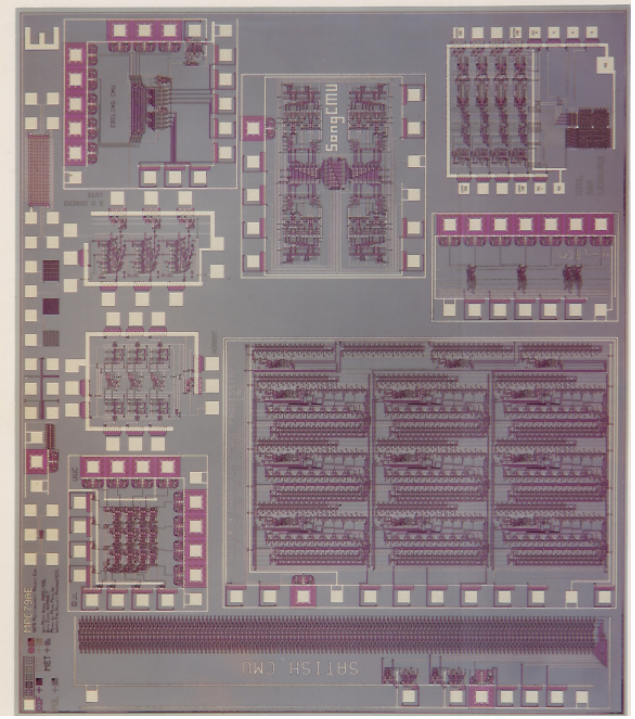
Pastilha multiprojeto

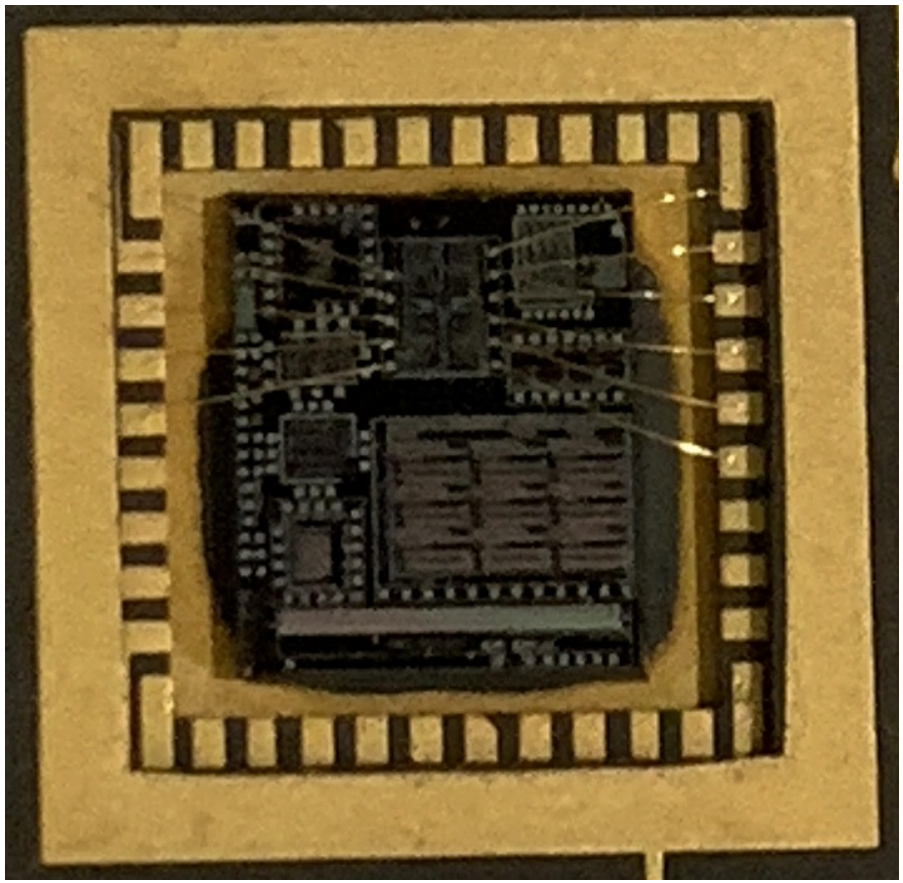


Pastilha multiprojeto



Pastilha multiprojeto





Array sistólico - um exemplo

- Vamos mostrar um exemplo de um array sistólico que multiplica duas matrizes.

  Clicar aqui para os slides (.pptx).

  Clicar aqui para um vídeo (mp4).

Array sistólico - a moda vai e volta

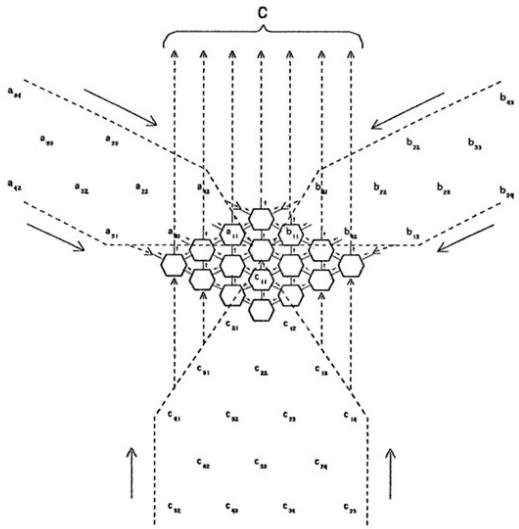
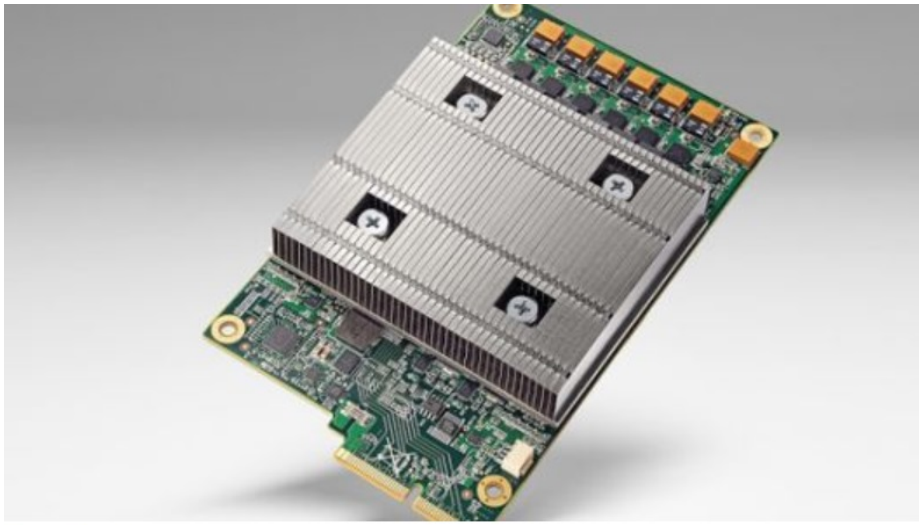


Figure 4-2: The hex-connected systolic array for the matrix multiplication problem in Figure 4-1.

- Proposto em 1978, array sistólico despertou enorme interesse na época.
- Mas com o tempo a moda passou e ficou latente durante quase trinta anos.
- Até que ressurge em 2016 pela Google TPU (Tensor Processing Unit).

Google TPU - Tensor Processing Unit 2016 - 2024

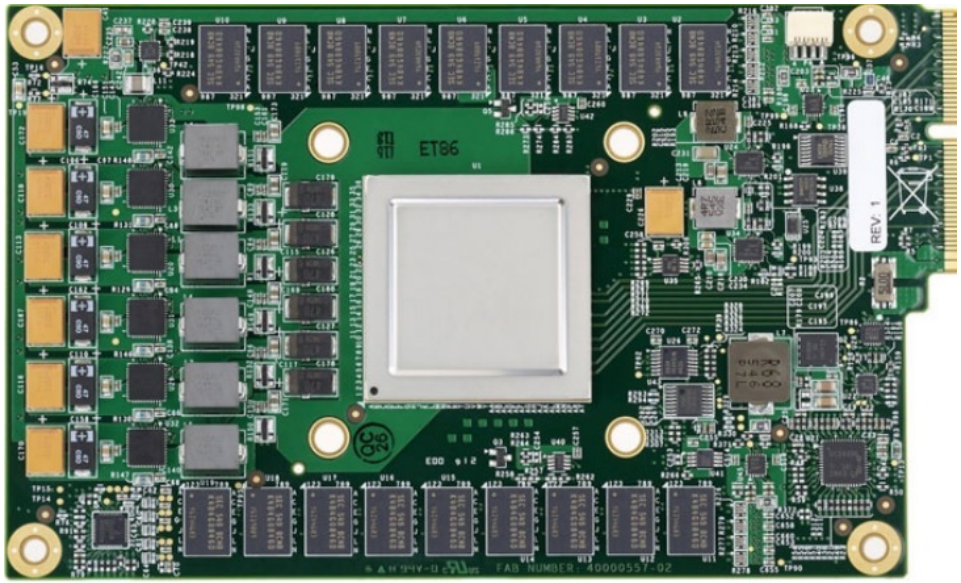


Source: Google

Google's tensor processing unit or TPU.

- Array sistólico ressurge na figura da Google TPU (Tensor Processing Unit) que é usado em Google Search, Google Street View, Google translate para acelerar as computações de redes neurais artificiais em aprendizado de máquina.

Google TPU - Tensor Processing Unit 2016



Google's first TPU

- Primeira geração TPU (2016): um 256 $\times$ 256 array sistólico que realiza multiplicação de matrizes de números inteiros de 8 bits, e operação de convolução.

Google TPU - Tensor Processing Unit 2017

Tensor Processing Units in a data center

– Google

- Segunda geração TPU (maio 2017): multiplicação de matrizes em ponto flutuante, com desempenho de 11,5 PetaFLOPS (PFLOPS).
- Quatro TPUs de 45 TFLOPS cada são arranjados em módulos de 4 cada, dando 180 TFLOPS. Então 64 módulos são montados em pods de 256 chips cada totalizando 11,5 PFLOPS.

An in-depth look at Google's first Tensor Processing Unit (TPU). Kaz Sato et al. May 12, 2017.

Google TPU - Tensor Processing Unit 2018 a 2024

- Terceira geração TPU (maio 2018): oito vezes mais rápido que TPU da segunda geração, com 1.024 chips por pod.
- Quarta geração TPU (maio 2021): duas vezes mais rápido que TPU da terceira geração, com 4.096 chips por pod.
- Quinta geração TPU (dezembro 2023): duas vezes mais rápido que TPU da quarta geração.
- Sexta geração TPU (anunciada em maio 2024 para final de 2024): Denominada de Trillium.

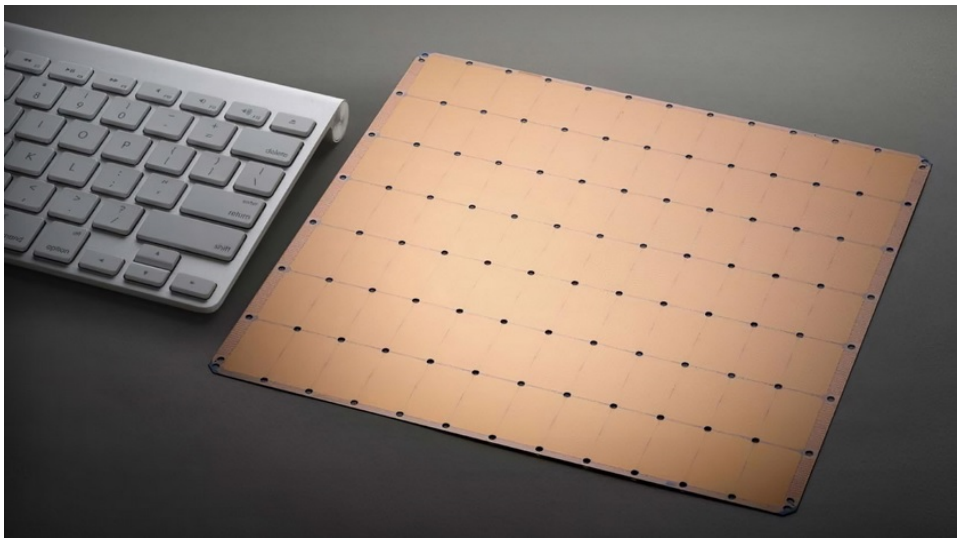
Google TPU - Tensor Processing Unit 2024

Sundar Pichai announces Google's sixth-generation TPU at the Google I/O developer conference.

- Sexta geração TPU (anunciada em maio 2024 para final de 2024): Denominada de Trillium, será 4,7 vezes mais rápido que TPU da quinta geração.
- Trillium tem até 256 TPUs em um pod, podendo chegar a centenas de pods.

Announcing Trillium, the sixth generation of Google Cloud TPU.

WSP - Wafer Scale Processing - 2,6 trilhão transistores



- WSP - Wafer Scale Processing: usar todo o wafer para uma CPU
- Cerebras WSP com 2,6 trilhão transistores e 850.000 *cores*.
- Tecnologia de 7 nm.

Próximo assunto: Como aumentar o desempenho do processador

- Arquiteturas X86 e ARM.
- Técnicas para aumentar o desempenho do processador
- Ao longo dos anos, várias técnicas foram criadas visando maior velocidade do processador.
- Em 2018 vulnerabilidades (Meltdown e Spectre) foram descobertas que exploram essas técnicas. (Vermos Meltdown e Spectre mais tarde, primeiro vamos ver as tais técnicas...)
- Não percam!