

# 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>

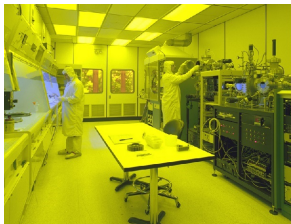
Baseado em parte em Mead and Conway - Introduction to VLSI Systems  
Esse assunto não cai em provas

# 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 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  $0,5 \mu\text{m}$  por  $\text{m}^3$ .
- Sala limpa ISO 1:  $\leq 12$  partículas de  $0,3 \mu\text{m}$  por  $\text{m}^3$ . 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

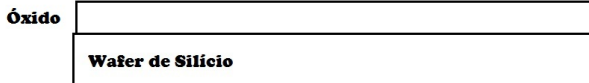


Source: PublicDomainPictures.net

- 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.

## **Wafer de Silício**

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

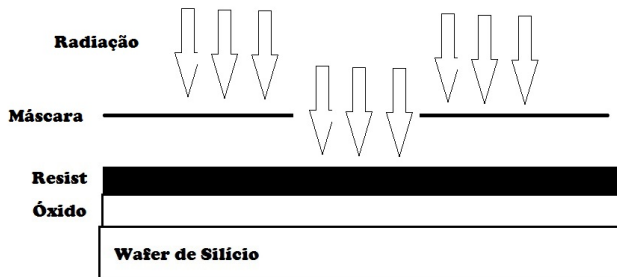


- Expor wafer de silício a oxigênio num forno de alta temperatura. Forma-se óxido  $SiO_2$  na superfície.

# Processo Básico

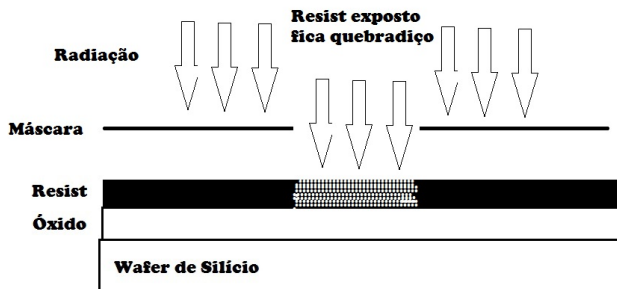


- Pintar com uma fina camada de material orgânico chamado “resist”. Secar e “assar” no forno.



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

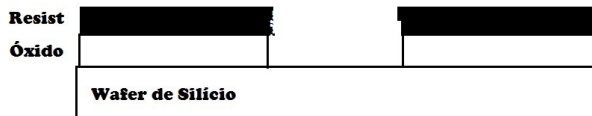




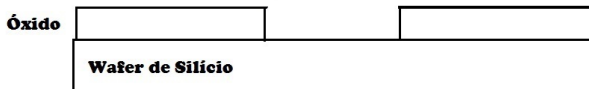
- Isso vai quebrar a estrutura de moléculas de parte (expostas) do resist.



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



- Usar ácido hidrofúrico que dissolve o óxido  $\text{SiO}_2$  mas não ataca o “resist”.



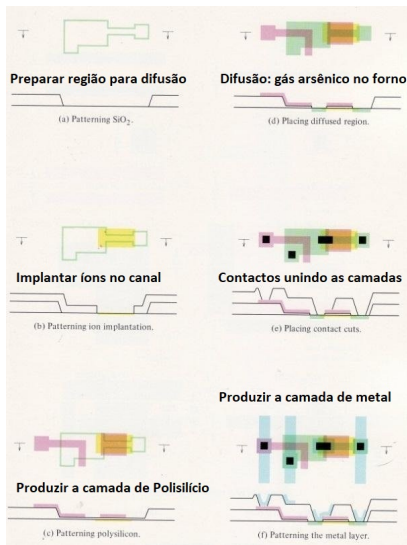
- Eliminar “resist” com solventes fortes ou ácidos. ○ processo básico produz a forma da máscara no chip.

# Processo Completo

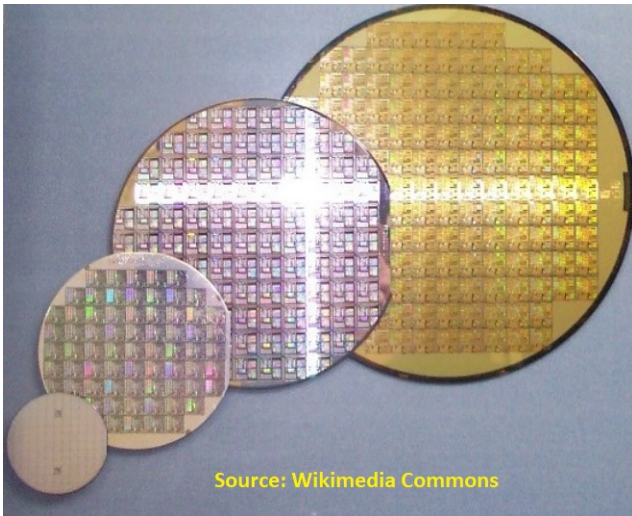
- O processo básico produz a forma da máscara no chip.
- Lentes de redução colocadas entre a máscara e o wafer reduzem a forma da máscara ao chegar no wafer.
- O processo básico é usado no processo completo para produzir as várias camadas (difusão, poli-silício, metal, etc.) na pastilha, conforme as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

# Processo Completo

Source: Mead and Conway - Introduction to VLSI Systems

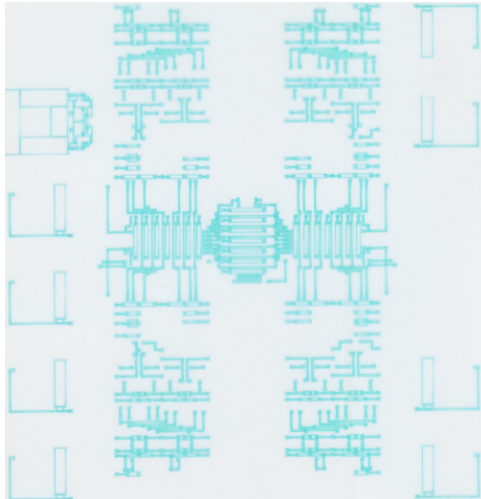


# Wafers de Silício



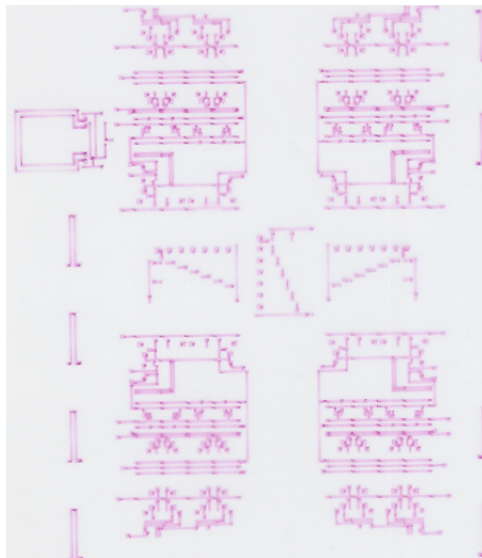
Source: Wikimedia Commons

# Máscara Difusão

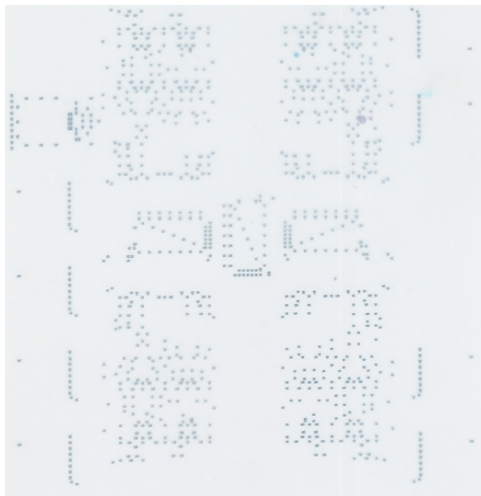




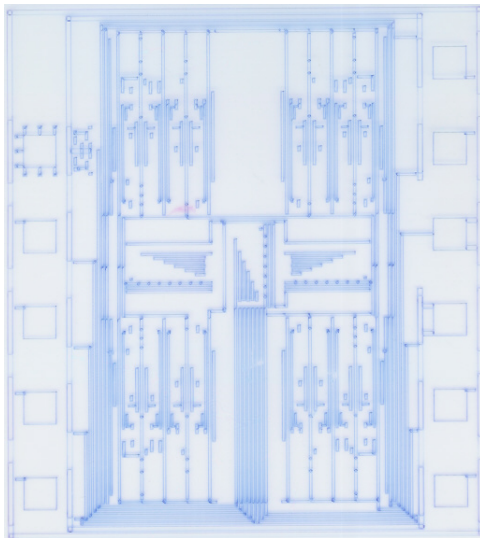
# Máscara Poli-silício



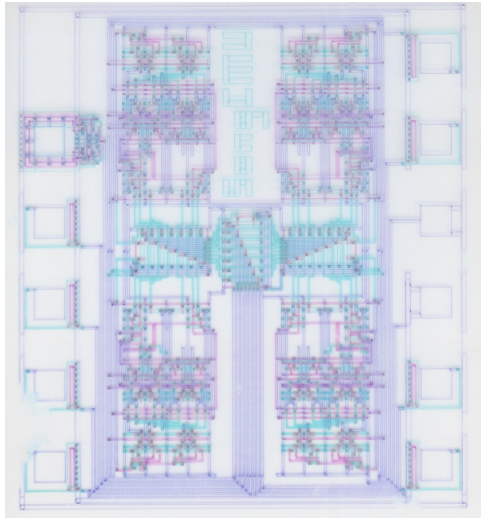
# Máscara Contatos



# Máscara Metal



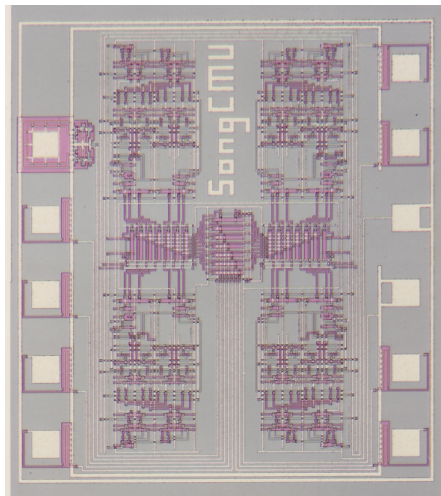
# Todas As Camadas Juntas



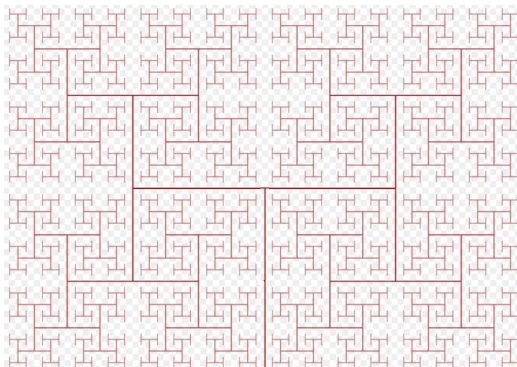
# Projeto de pastilha customizada para aplicação específica


- 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). 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 conjunto de células (elementos de processamento) simples interconectadas de uma forma regular no plano.

# Projeto busca por árvore binária

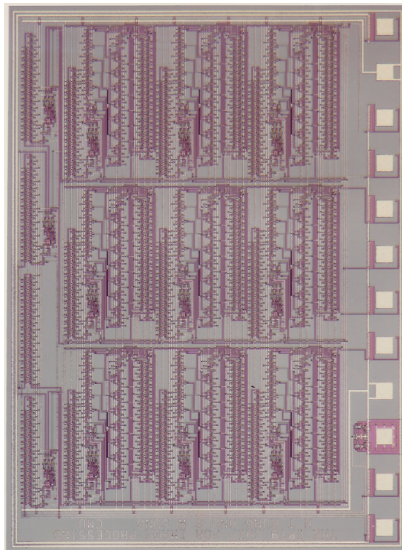


# Disposição-H de uma árvore binária no plano



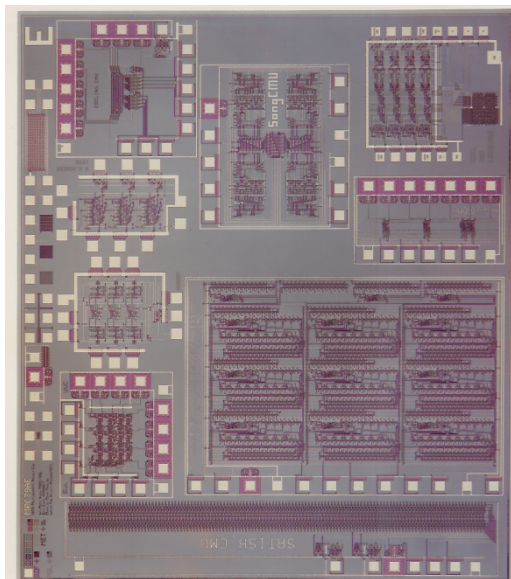
- Acima mostramos uma árvore binária em disposição-H (nome devido à forma H que aparece no desenho) que melhor utiliza o espaço.
- Quantos nós tem essa árvore acima? (Tente desenhá-la na forma usual de representar uma árvore binária (i.e.: ) no mesmo espaço acima :-)

# Projeto convolução

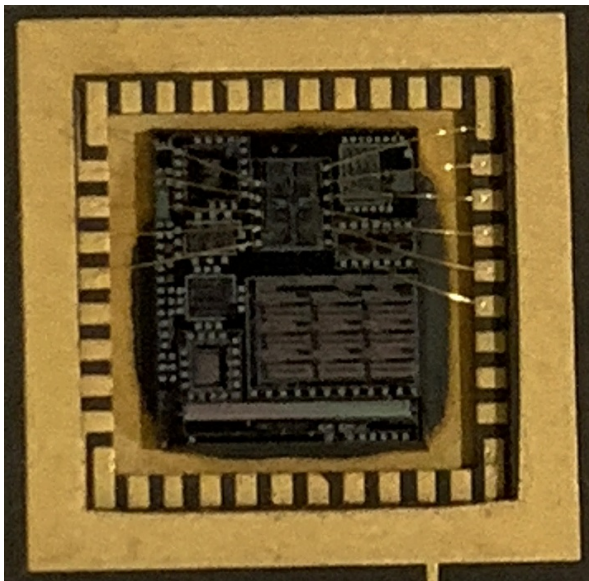




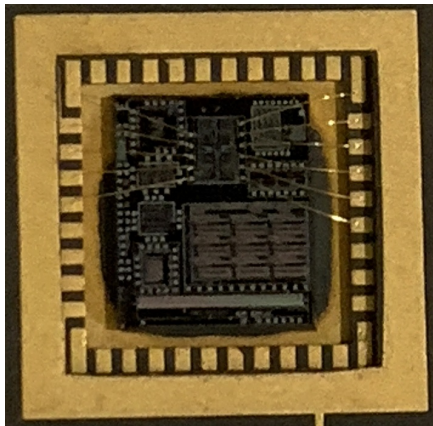
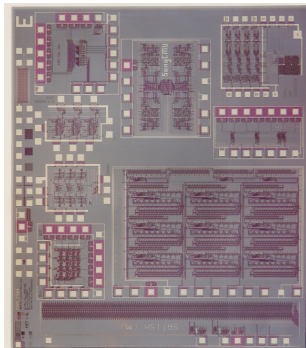
# Pastilha multiprojeto



# Pastilha multiprojeto



# Pastilha multiprojeto



# Array sistólico - um exemplo

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

[Clicar aqui para ver o exemplo \(mp4\).](#)

# Array sistólico - a moda vai e volta

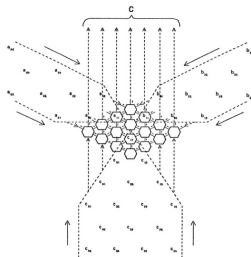


Figure 4-2: The hex-connected systolic array for the matrix multiplication problem in Figure 4-1.

- 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 - 2018

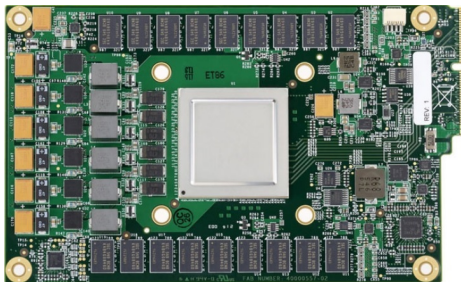


Source: Google

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** 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**.

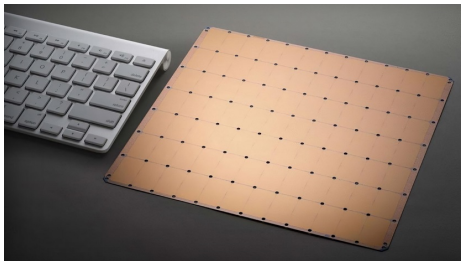
# Google TPU - Tensor Processing Unit 2018

- Segunda geração TPU (maio 2017): **multiplicação de matrizes** em ponto flutuante, com desempenho de 11,5 PetaFLOPS, usada no treinamento e inferência em **redes neurais** para **aprendizado de máquina**.
- Terceira geração TPU (maio 2018): oito vezes mais rápido que TPU da segunda geração.

An in-depth look at Google's first Tensor Processing Unit (TPU). Kaz Sato (Staff Developer Advocate, Google Cloud), Cliff Young (Software Engineer, Google Brain), David Patterson (Distinguished Engineer, Google Brain) May 12, 2017.



# 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.

# Principais fabricantes de chips VLSI

Hoje existem 3 fabricantes no mundo capazes de produzir chips com a tecnologia de 7 nm. ([Clicar aqui para a reportagem completa.](#))

- Taiwan Semiconductor Manufacturing Company (TSMC)

Para um vídeo sobre esse fabricante, ver:

[Inside The World's Largest Semiconductor Factory - BBC](#)  
(4:17 minutos)

- Samsung
- Intel

A previsão é que em 2024 será possível produzir chips com a tecnologia de 5 nm.

# Próximo assunto: Como aumentar o desempenho do processador

- Próximo assunto: 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!