Superconductor Computing for Neural Networks

题外话

电阻为零的超导微处理器问世(2021-01-18)

> "这个新的微处理器原型称为MANA（单绝热集成体系结构），是世界上第一个绝热超导体微处理器。它由超导铌组成，并依赖于称为绝热量子通量参量电子（AQFP）的硬件组件。每个AQFP由几个快速作用的约瑟夫森结开关组成，这些结开关只需很少的能量即可支持超导体电子设备。MANA微处理器总共由2万多个约瑟夫森结（或1万多个AQFP）组成。"

> The microprocessor mainly consists of an on-chip instruction memory, two data registers, an instruction decoder, an 8-bit bit-parallel arithmetic logic unit, and a program counter. The microprocessor contains 7702 JJs (based on the Open Dataset of CONNECT Cell Library for AIST ADP2) without considering splitters, Josephson transmission lines, and passive transmission lines.

Dennard缩放比例定律

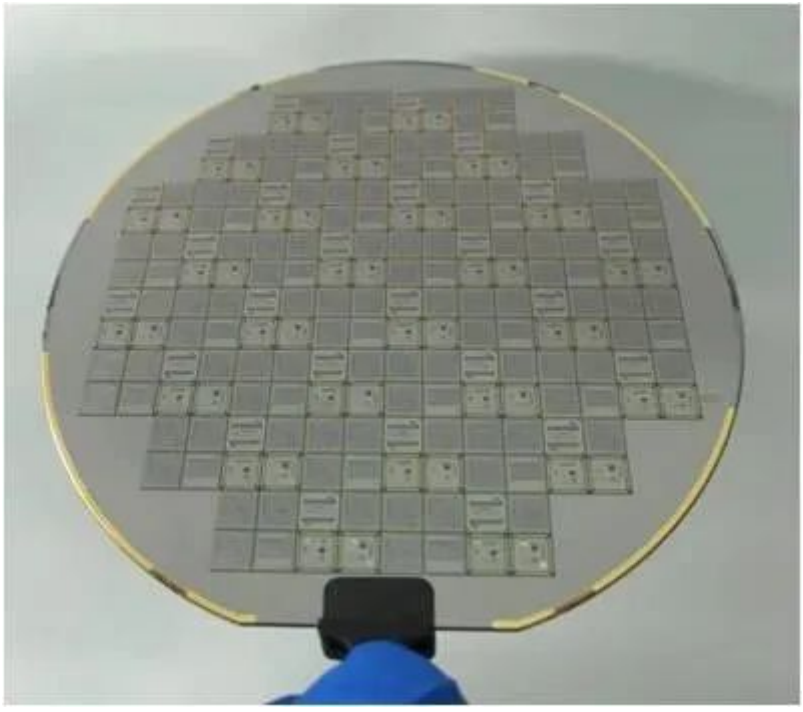
Dennard缩放比例定律是基于1974年Robert H. Dennard参与完成的一篇论文而提出的一个定律。

Dennard缩放比例定律（Dennard scaling）表明，随着晶体管变得越来越小，它们的功率密度保持不变，因此功率的使用与面积成比例；电压和电流的规模与长度成比例

> 自2005-2007年前后，Dennard缩放比例定律似乎已经失效。截至2016年，集成电路中的晶体管数量仍在跟随“摩尔定律”增加，但由此带来的性能改善却更为缓慢。这种情况的主要原因是在芯片尺寸不变，晶体管数量变多的情况下，电流泄漏会带来更大的挑战，也会导致芯片升温，从而造成热失控的威胁，从而进一步增加能源成本.

任洁：让超算更快、更小、更节能——超导数字集成电路|2022年·第10期(2023-01-10)

- 超导集成电路集成在超导晶圆



> 再者，这里的“数字”，是说它依旧是进行0和1的运算，只不过，这里的0和1信号不再是用电压的高和低来代表，而是用单磁通量子的脉冲信号在超导环路中的有和没有来代表0和1。

- 超导集成电路的三个特别的要素: 材料, 器件, 逻辑

首先是材料的发现。超导这个现象是1911年由H. K. Onnes 首先发现的，他把氦气变成液氦，在4.2K的温区，发觉液氦里面放入的金属完全没有电阻了。这就是我们低能耗的一个最基本的来源之一。此外，超导特有的现象——磁通量子化也与超导集成电路息息相关。磁通量子化的意思就是说，超导环路内有且仅有整数倍的磁通量子，这样，我们就可以选择整数倍为零或一来代表逻辑的0和1

第二点，开关器件。超导集成电路的开关是一个叫约瑟夫森结的器件，它像三明治一样，是两块超导体夹以某种很薄的势垒层

在约瑟夫森效应被发现不久，美国以及日本的很多计算的先锋，都尝试利用约瑟夫森结来进行计算机的研发，它第一代的逻辑其实就跟CMOS一样，用电压的高和低来代表数字的0和1，并产出了很多芯片。但随着CMOS的迅速发展，约瑟夫森结作为高低电压开关的速度并没有压倒性的优势，再加上超导器件对于环境要求严苛，人们没有选择使用它。

1985年，三位莫斯科国立大学的科学家提出了一种新的逻辑形式——SFQ逻辑门。它虽然还是用约瑟夫森结作为开关，但是它的数据形式从电压电平逻辑变成了电压脉冲的逻辑，这样怎么进行计算呢？还是以“与”门为例，CMOS的“与”门，是通过判断两个输入是高还是低来决定输出是高还是低，但是SFQ的“与”门，比它多加了一个时钟信号，这是因为两个脉冲很难同时到达，所以我们允许它有一点时间间隔。我们在判断逻辑输出的时候，当某一个时钟到来，我们就通过观察距离上一个时钟的时钟周期内，是否又来了一个脉冲，如果有，“与”门就输出1，如果没有，那就输出0。在SFQ逻辑里，其他的门都是类似的。

https://mozi.ustc.edu.cn/detail/999

Basic Concepts

《超导电子技术》

1. 迈斯纳效应(完全抗磁性)

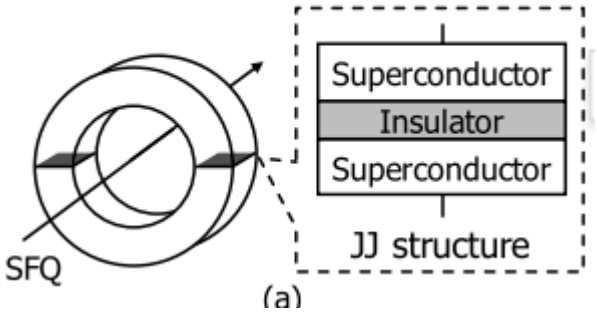
- 不管加磁场的次序如何, 超导体内磁场感应强度总是等于0, 它与施加磁场的历史无关

- 在理想导体中, 不能存在电场, 也就是说理想导体中的磁感应强度不随时间变化
- 超导体和理想导体的最大区别就是迈斯纳效应

2. RSFQ-based Circuit 的基本原理

- 最早使用约瑟夫森结的两种电压状态来表示逻辑1和0, 超导态作为逻辑0, 常态作为逻辑1, 这种结构称为**栓锁结构**
- 由于**回滞现象**的存在, 只有当电流趋近于0时, JJ才能回复到超导态, 需要施加时钟交变偏置电流, 但随着时钟交变偏置电流频率的提高会使电流过冲使超导体进入常态, 由此对clk频率的提高造成限制
- RSFQ技术（快单磁通量子技术）利用超导电路在特定时隙中是否含有单磁通量子电压脉冲来表示逻辑的“0”和“1”。

超导环为什么被设计为环状结构?



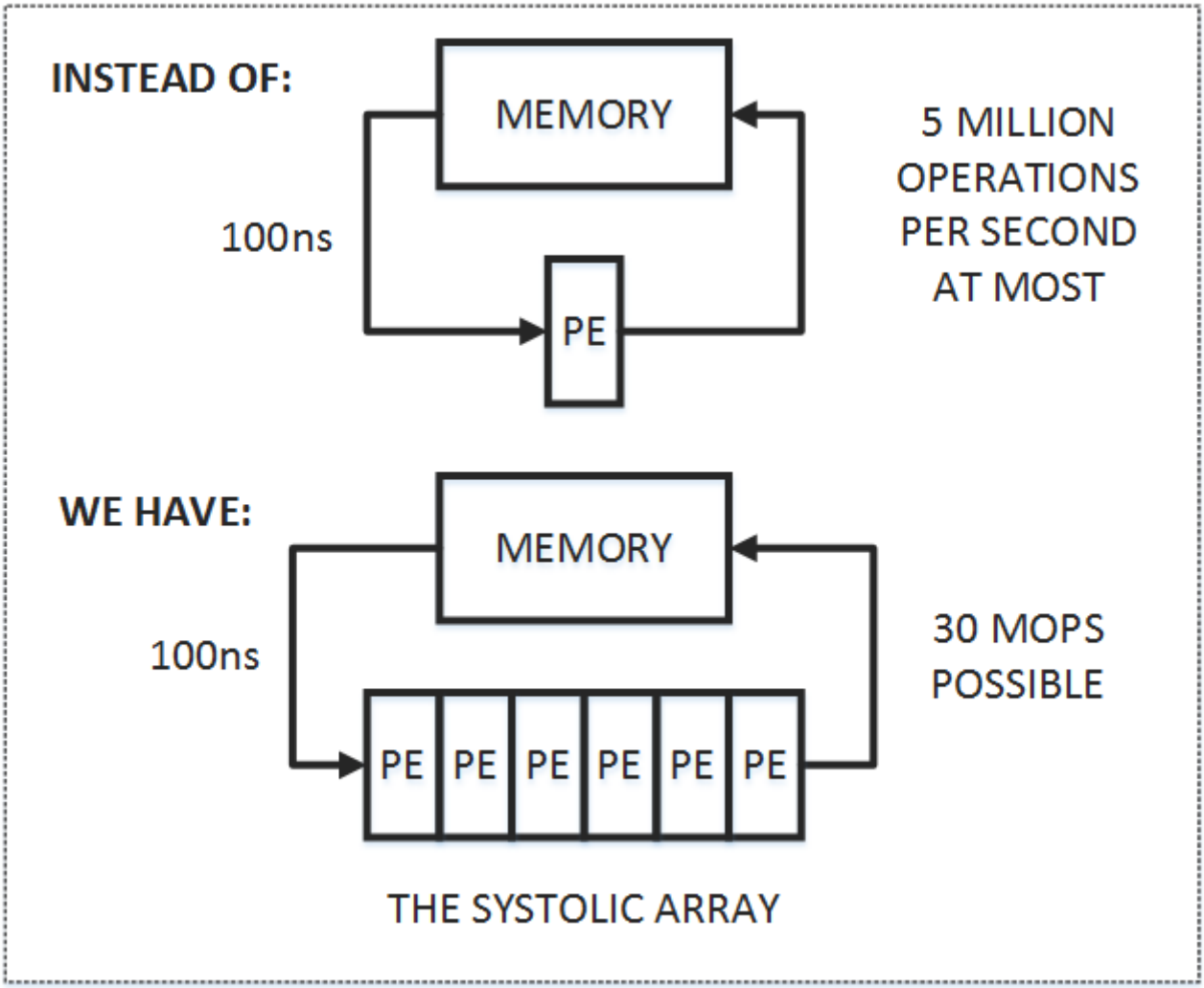
- 因为环形的几何结构在电路中引入了磁通量量子现象, 并显示了一些特殊的超导性质, 例如量子磁通量量化效应和环电流
- 根据约瑟夫森效应, 环电流必须是一个磁通量量子的整数倍, 这使得环状结构成为研究约瑟夫森效应和磁通量量子化的理想选择

Superconductor ring磁通量子化(flux quantization)

- 在垂直超导环表面的方向上有一个磁场B, 当温度降低到临界温度Tc时, 环就变成超导体.
- 由于迈斯纳效应, 磁场被排斥在超导环之外, 但是一部分磁通量仍陷于环所包围的空间. **若此时移去外加磁场, 则超导环所包围的磁通不会随之消失, 超导环表面环流的存在维持磁通**

Systolic array

一种特殊的并行计算结构; 主要由一组紧密排列的处理单元组成, 受到心脏的搏动现象的启发, 类似地, systolic array中的数据在处理单元之间通过紧密的同步和协作方式流动, 实现高效的并行计算.



> 图中上半部分是传统的计算系统的模型。一个处理单元（PE）从存储器（memory）读取数据，进行处理，然后再写回到存储器。这个系统的最大问题是：数据存取的速度往往大大低于数据处理的速度。因此，整个系统的处理能力（MOPS，每秒完成的操作）很大程度受限于访存的能力。而脉动架构用了一个很简单的方法：让数据尽量在处理单元中多流动一会儿。

> 第一个数据首先进入第一个PE，经过处理以后被传递到下一个PE，同时第二个数据进入第一个PE。以此类推，当第一个数据到达最后一个PE，它已经被处理了多次。所以，脉动架构实际上是多次重用了输入数据。因此，**它可以在消耗较小的memory带宽的情况下实现较高的运算吞吐率**。当然，脉动架构还有其它一些好处，比如模块化的设计容易扩展，简单和规则的数据和控制流程，使用简单并且均匀的单元（cell），避免了全局广播和扇入（fan-in），以及快速的响应时间（可能？）等等。

Content

Background on SFQ logic Technology

The working principle of SFQ logic gates with an SFQ-based delay flip flop

1. First, when the input pulse enters the ring, it is stored in the ring as an SFQ by switching JJ1.
2. Next, by receiving a clock pulse, JJ2 is activated, and the stored SFQ is transferred to the output as a voltage pulse .
3. In this manner, SFQ gates can represent the logical value “1” (or “0”) by the existence (or absence) of the stored SFQ between the two adjacent clock pulses.

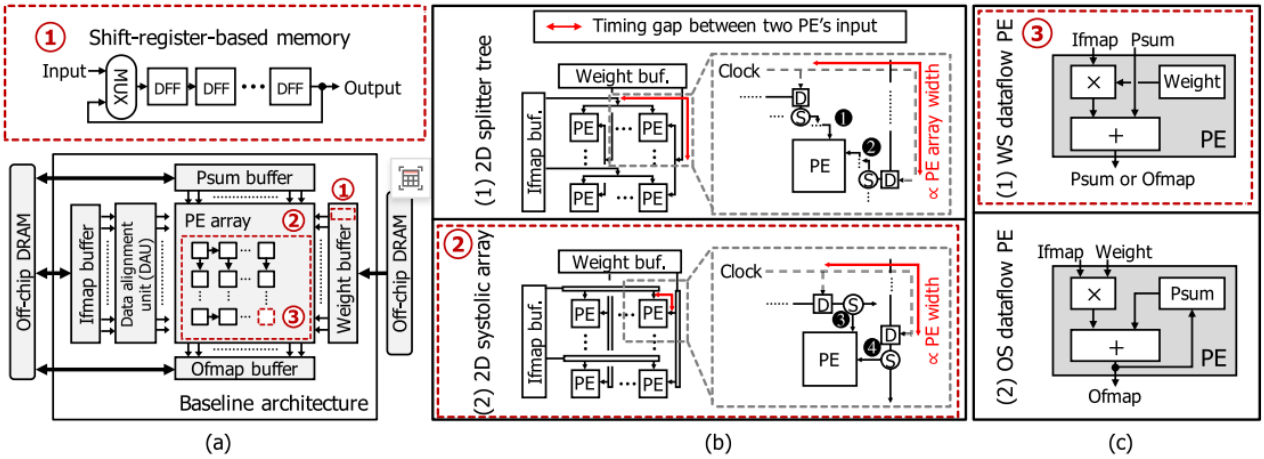


FIGURE 2. Our baseline SFQ-based NPU with each microarchitecture unit's design alternatives. (a) Overview of our baseline architecture. (b) Network designs. (c) PE designs.

- SFQ technology favors simple control flows due to its deeply pipelined nature.

Design of Max Pooling Operation Circuit for Binarized Neural Networks Using Single-Flux-Quantum Circuit

Basic Concepts

Binary Neural Network(BNN)

Max pooling operation circuit

MPOC是一种电路或硬件实现, 用于执行神经网络中的最大池化操作

- 最大池化操作是深度学习中常用的一种操作, 通常用于减少数据的维度并提取关键特征
- 最大池化操作的基本思想是在输入数据的局部区域中选择最大值作为输出, 以减少数据的空间维度, 并且强调数局部特征的重要性. 最大池化通常通过滑动一个固定大小的窗口来完成, 将窗口内的最大值作为输出, 并将窗口按固定的步幅在输入数据上滑动以覆盖整个输入区域.

MPOC的作用是在硬件层面上实现这一最大池化操作, 以便在嵌入式系统, 专用硬件加速器或其他需要高效神经网络推理的应用中执行

- 窗口指的是在输入数据上滑动的固定大小的区域或矩形, 它在输入数据上按照一定的步幅stride从左到右, 从上往下滑动, 覆盖输入数据的不同部分
- 窗口的大小由两个参数决定:

1. 窗口大小: 通常以像素为单位
2. 步幅sride:步幅定义了在输入数据上移动窗口的距离, 通常以像素为单位

- 最大池化操作的过程:

1. 将窗口的起始位置放在输入数据的左上角
2. 在窗口内部找到最大值
3. 将该最大值作为输出的一部分

4. 将窗口按照步幅的设置在输入数据上滑动到下一个位置
5. 重复步骤2-4, 直到窗口覆盖了整个输入数据, 产生了最大池化后的输出数据

- 通过使用窗口大小和步幅参数, 最大池化操作能够在保留重要特征的同时减小输入数据的维度, 有助于提取关键信息并降低模型的计算复杂度

channel

> channel通常指的是数据张量中的一个维度, 用于表示不同的特征或信息. 图像处理中用三通道来表示彩色图像的红色, 绿色, 蓝色信息

在文中的"1-channel 3*3 convolutional layer"意味着这是一个卷积神经网络的卷积层, 该层具有一个输入通道; 这意味着该卷积层接受单通道的输入数据, 通常用于处理灰度图像或其他单通道数据

feedback loop

- 系统内的输出被用作输入，以此循环，继续产生输出，这个过程称为反馈环。

反馈回路的最用是帮助系统实现稳定性, 准确性和自适应性; 通过反馈, 系统可以纠正误差, 调节参数, 调整行为或控制输出, 以使系统的行为更贴合预期目标.

Contents

Max Pooling Hardware Design

Max Pooling

> The pooling layer is an important part of BNN, which can effectively avoid overfitting and reduce the amount of computation by subsampling.

overfitting: The CNN may fit the training set very well, but fail to generalize to new examples.

对于传统的comparator, 一次只能比较两个数据, 不能输出最大值. 需要很多移位寄存器来保存数据, 获得最大值. 并且很多comparator需要用一个树结构练级额, 让电路很难按比例放大.

- 传统的比较器的操作, 就是二叉树的排序算法.
- 为什么说树状结构会使得电路难以被放大? 比较器的输出通常需要进一步处理或放大, 以用于后续的电路. 然而树状结构会增加电路的复杂性和面积, 同时增加了功耗和成本
- 最大池化的操作中的移动窗口问题是一个很经典的算法题, 但我很久没写了...

Multiple Data Comparator

用传统的SFQ numercial comparator设计MPOC有两个主要的问题:

1. 只有numerical value才能被比较, 不能输出最大值, 所以就需要很多很多的移位寄存器(SRs)来存储数据, 多路复用器MUX来输出比较后的最大值
2. 由于反馈循环对SFQ circuit不适用, 那么多数据的比较就只能使用树状结构来连接多个comparator, 使得电路难以放大.

MPOC Hardware Design

每个clk都会产生一个当前最大值, 所以就会有很多个多余的数据, 因此在Multiple data comparator后面连接一个NDRO(Non-destructive read-out)

- 为什么要用6-bit SR来作为NDRO的输入时钟信号? the pipelined stages of 5-bit comparator is 6.

Simulation

> We compare the performance with MPOC using proposed multiple data comparator with MPOC using feedback loops and tree structures.

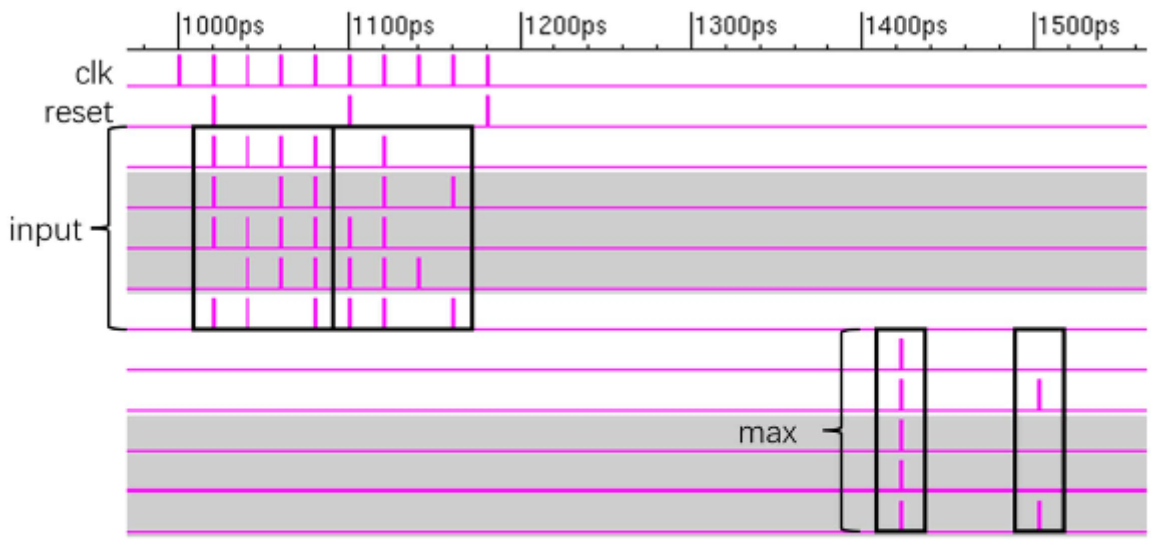


Fig. 6. Simulation results of the MPOC with high frequency (50 GHz).The designed bias voltage is 2.5 mV. For example in the first four clock cycles, We input "11101(-3)," "10111(-9)," "11110(-2)," "11111(-1)" to obtain the maximum value "11111(-1)" as output. Here MSB of input and max are sign bits.

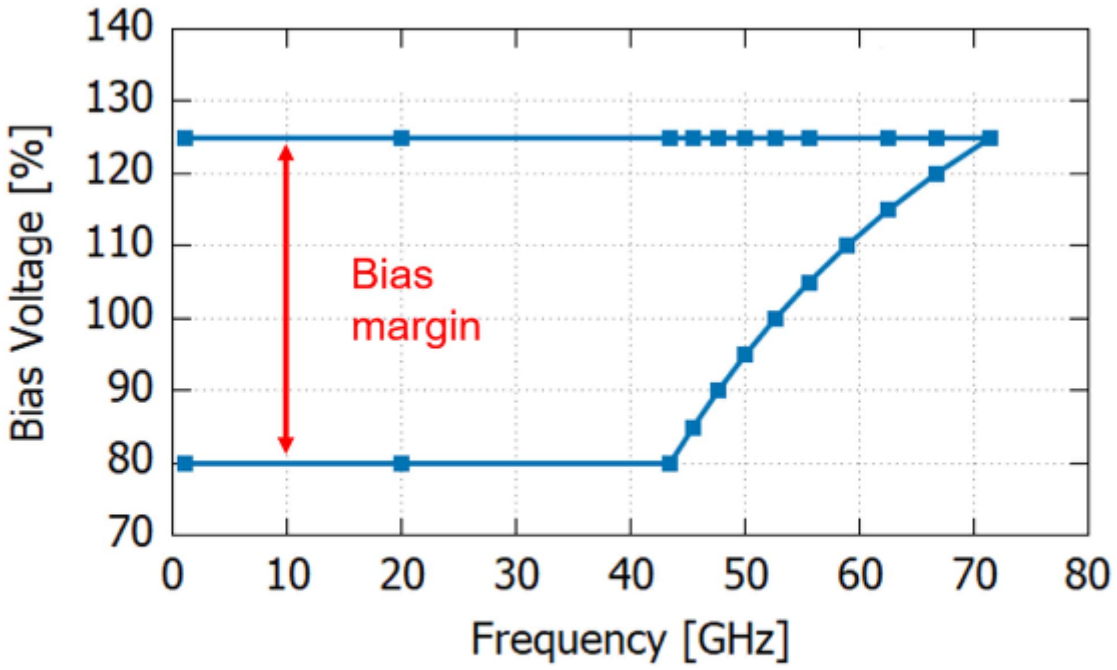


Fig. 7. Simulation results of the normalized bias voltage margin. The designed bias voltage is 2.5 mV.

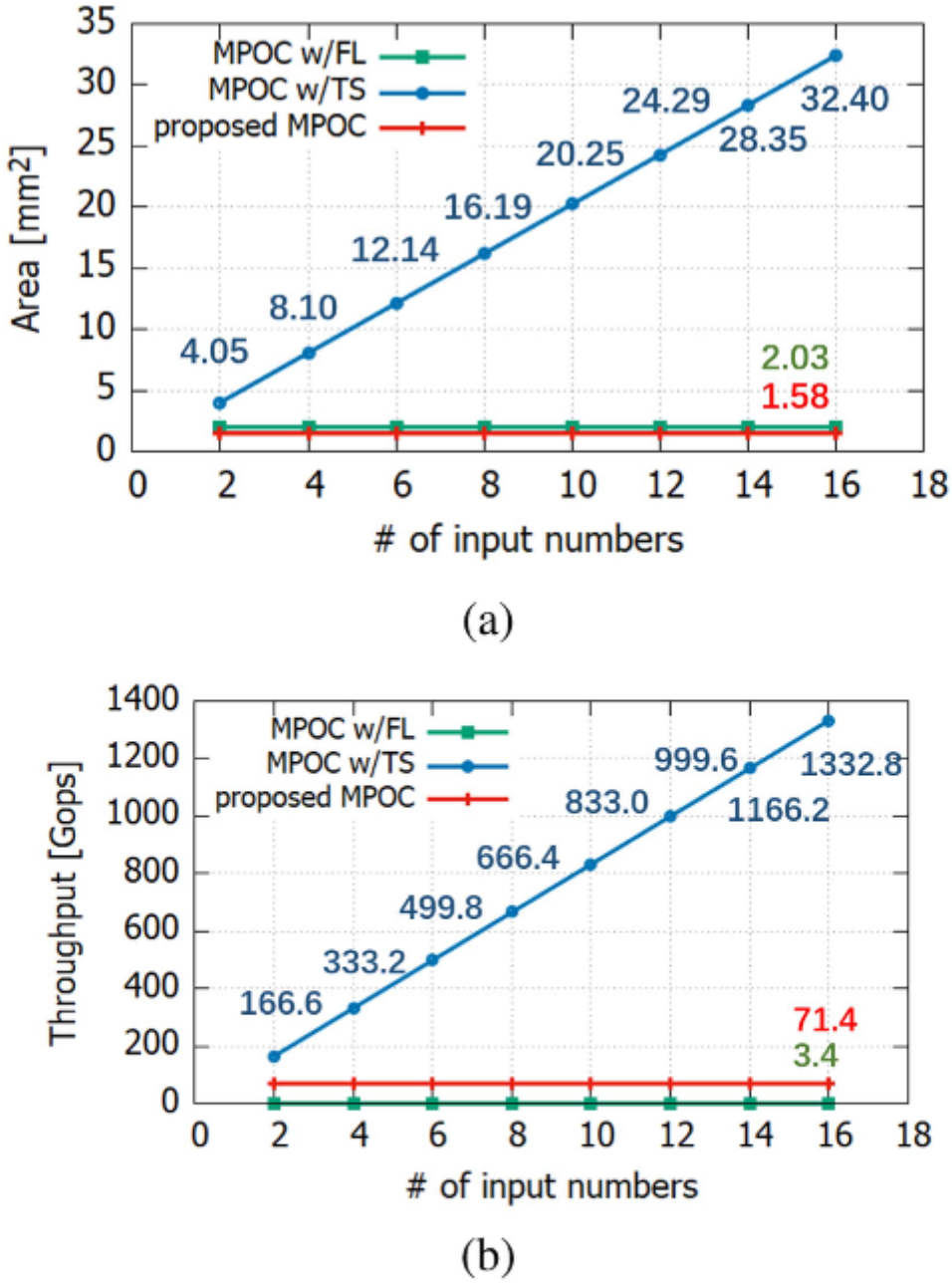


Fig. 8. Comparison of (a) the circuit area and (b) the throughput of the MPOC using the proposed comparator and conventional comparator as the number of inputs increase. conventional comparator w / TS is conventional comparator with tree structures and conventional comparator w / FL is conventional comparator with feedback loops.

References

脉动阵列 - 因Google TPU获得新生 - 唐杉的文章 - 知乎 https://zhuanlan.zhihu.com/p/26522315