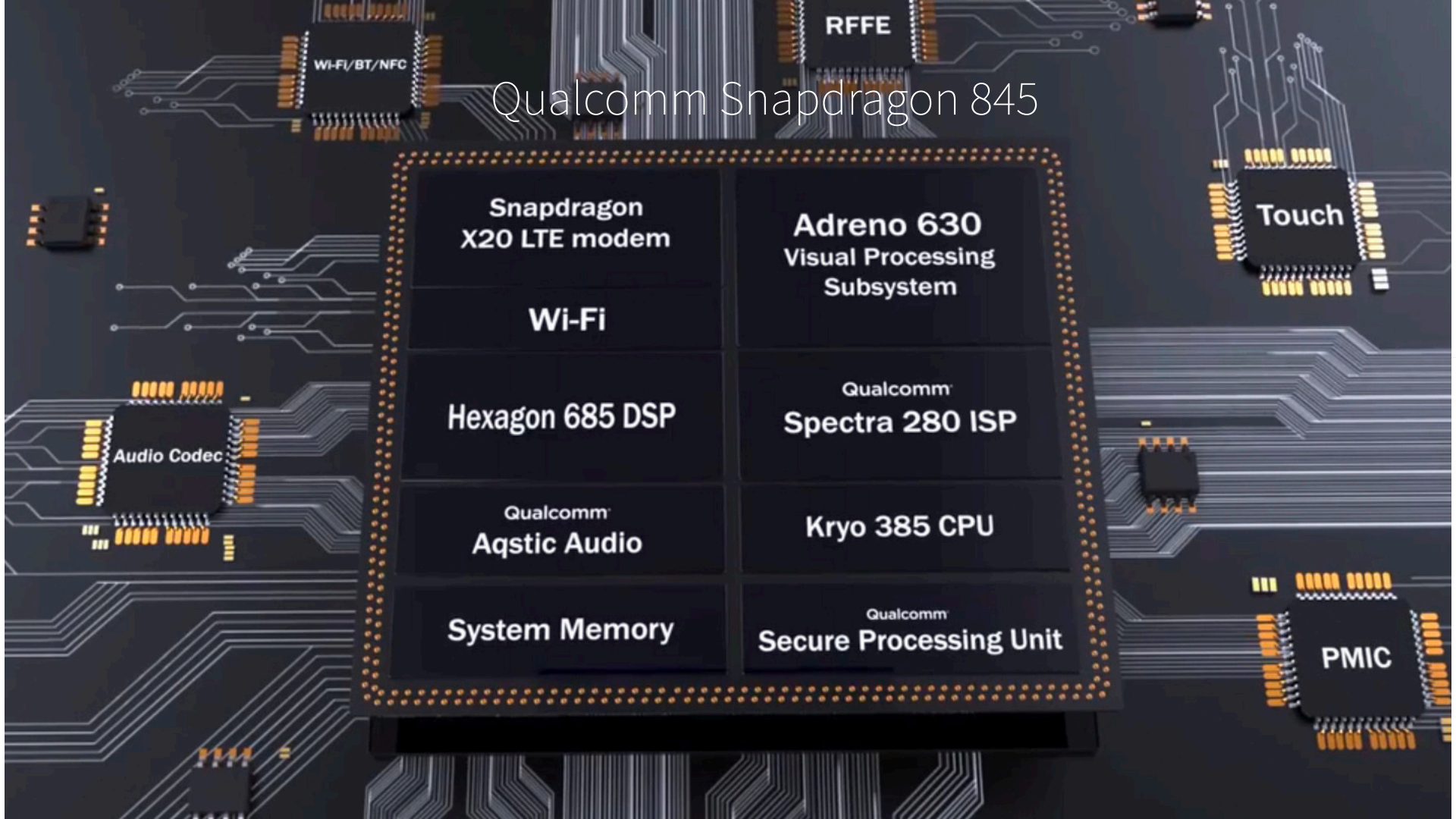
更多的芯片


Qualcomm Snapdragon 845
Wi-Fi/BT/NFC
RFFE
Touch
Audio Codec
PMIC
Snapdragon
X20 LTE modem
Wi-Fi
Adreno 630
Visual Processing
Subsystem
Hexagon 685 DSP
Qualcomm
Spectra 280 ISP
Qualcomm
Aqstic Audio
Kryo 385 CPU
System Memory
Qualcomm
Secure Processing Unit

DSP：数字信号处理

- 为数字信号处理算法设计：点积，卷积，FFT
- 低功耗、高性能
  - 比移动GPU快5x，功耗更低
- VLIW: Very long instruction word
  - 一条指令计算上百次乘累加
- 编程和调试困难
- 编译器质量良莠不齐

可编程阵列 (FPGA)

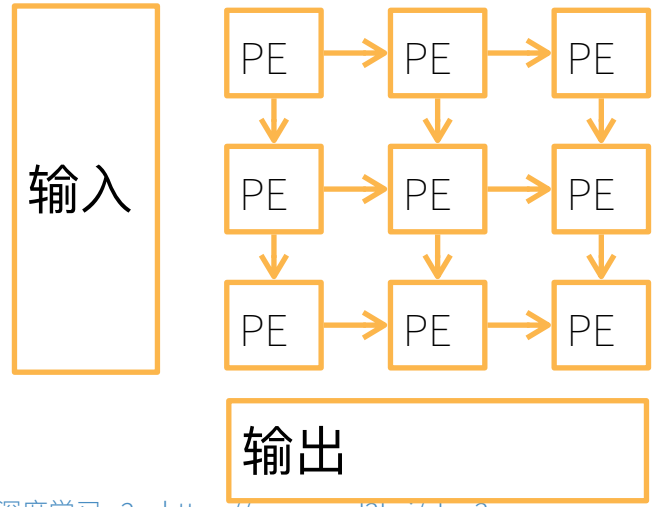
- 有大量可以编程逻辑单元和可配置的连接
- 可以配置成计算复杂函数
  - 编程语言：VHDL，Verilog
- 通常比通用硬件更高效
- 工具链质量良莠不齐
- 一次“编译”需要数小时

AI ASIC

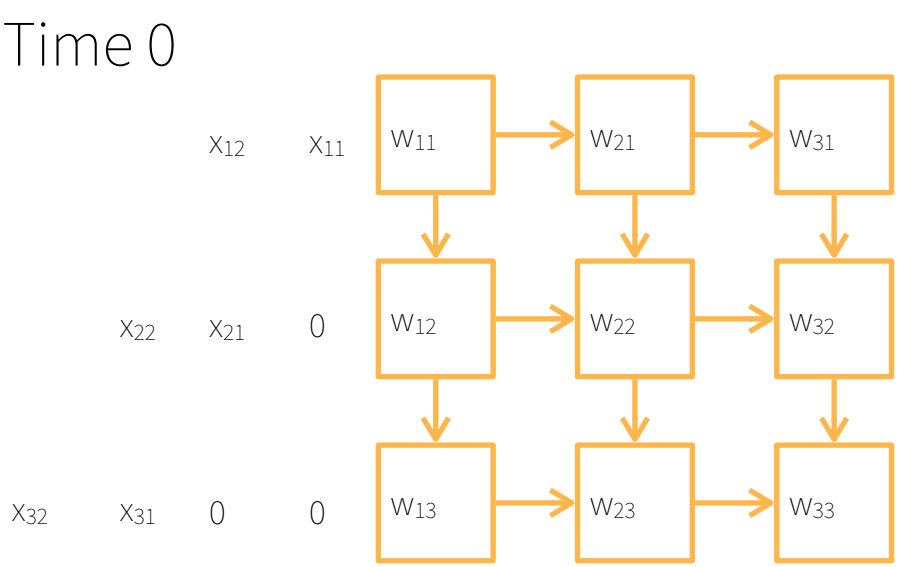
- 深度学习的热门领域
  - 大公司都在造自己的芯片 (Intel, Qualcomm, Google, Amazon, Facebook, ...)
- Google TPU 是标志性芯片
  - 能够媲美Nvidia GPU性能
  - 在Google大量部署
  - 核心是 systolic array

Systolic Array

- 计算单元（PE）阵列
- 特别适合做矩阵乘法
- 设计和制造相对简单



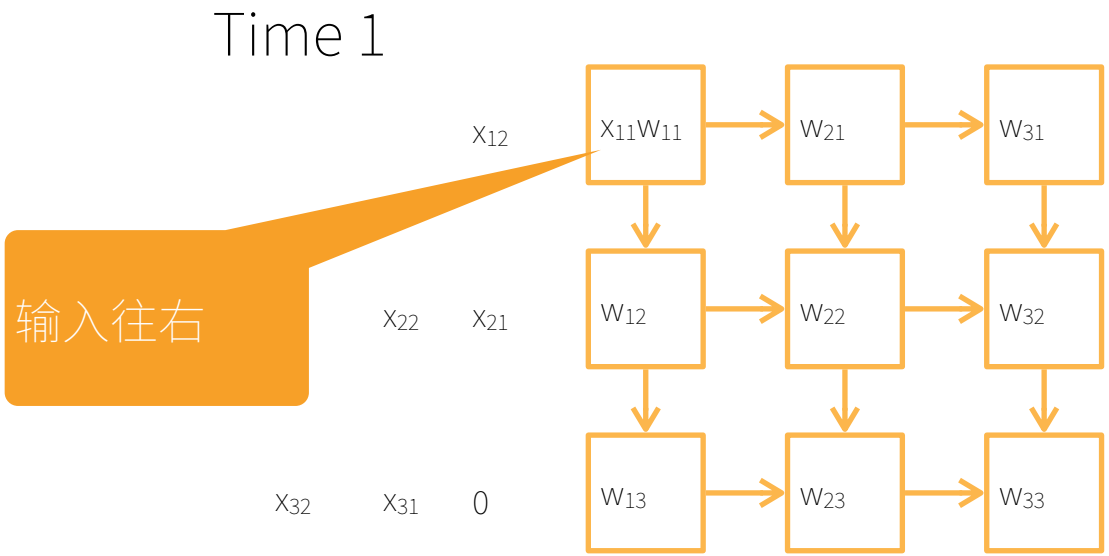
Systolic Array的矩阵乘法



$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

3x2 3x3 3x2

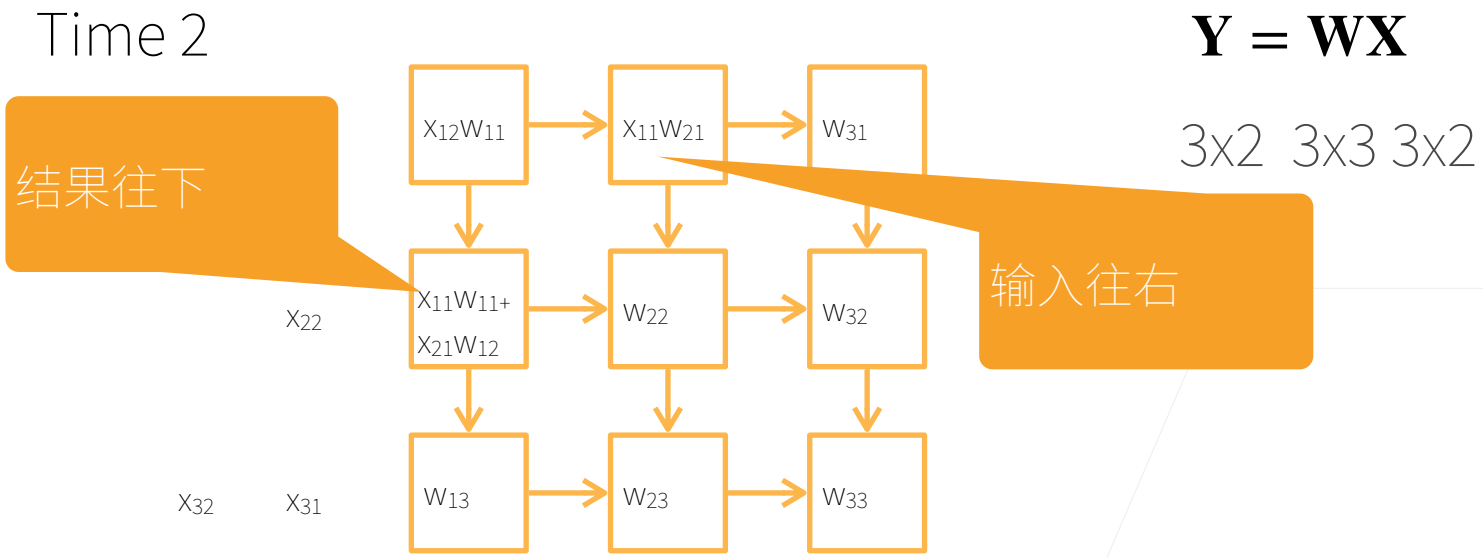
Systolic Array的矩阵乘法



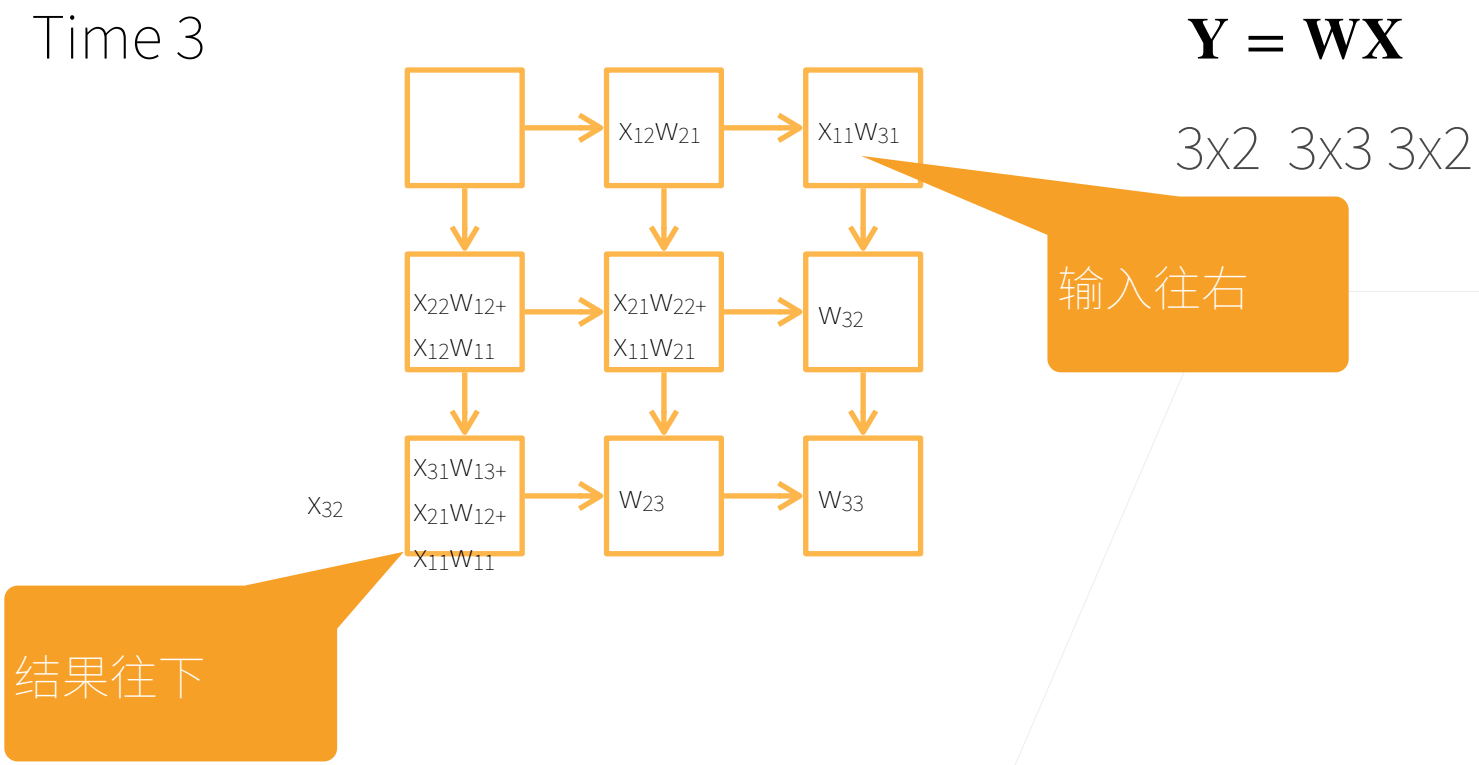
$$\mathbf{Y} = \mathbf{WX}$$

3x2 3x3 3x2

Matrix Multiplication with Systolic Array

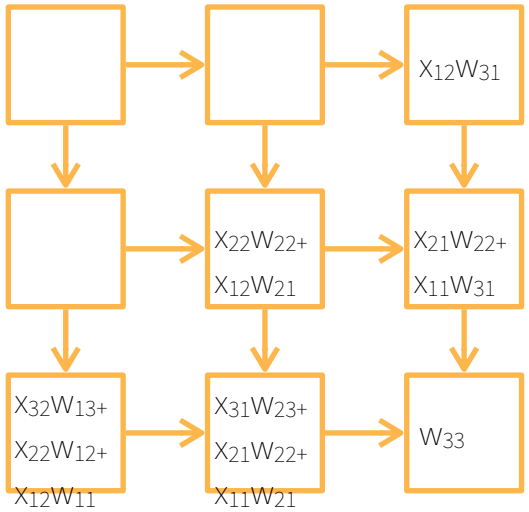


Matrix Multiplication with Systolic Array



Matrix Multiplication with Systolic Array

Time 4



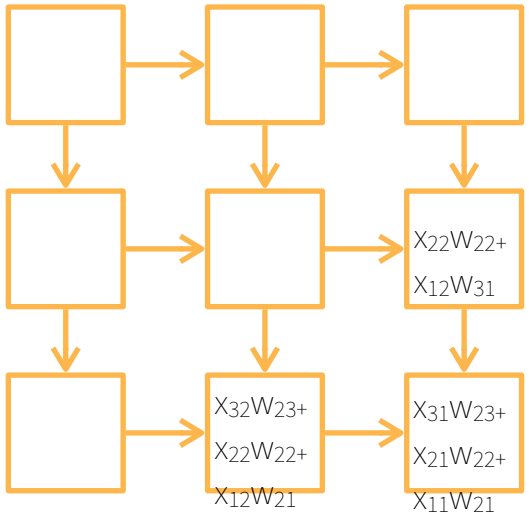
y_{11}

$$\mathbf{Y} = \mathbf{WX}$$

3x2 3x3 3x2

Matrix Multiplication with Systolic Array

Time 5

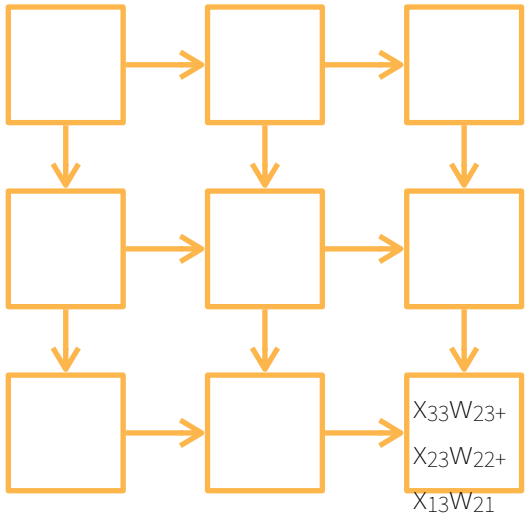


$$\mathbf{Y} = \mathbf{WX}$$

3x2 3x3 3x2

Matrix Multiplication with Systolic Array

Time 6

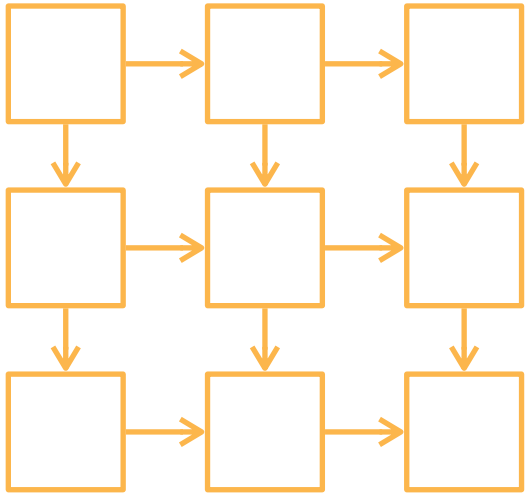


$$\mathbf{Y} = \mathbf{WX}$$

3x2 3x3 3x2

Matrix Multiplication with Systolic Array

Time 7



y_{32}

y_{22} y_{31}

y_{12} y_{21}

y_{11}

$$\mathbf{Y} = \mathbf{WX}$$

3x2 3x3 3x2

Systolic Array

- 对于一般的矩阵乘法，通过切开和填充来匹配SA大小
- 批量输入来降低延时
- 通常有其他硬件单元来处理别的 NN 操作子，例如激活层

总结


灵活性，易用性
intel
CORE i7
7th Gen
TITAN X
Hexagon 685 DSP
Xilinx®
FPGA
性能和功耗