主要内容

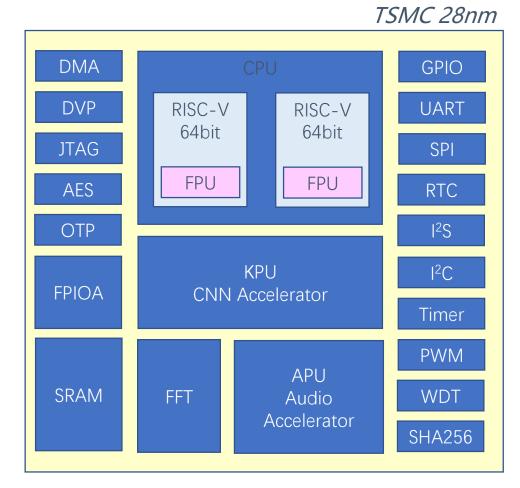


- 功能&接口
- 硬件设计
- KPU限制
- 模型转换
- 产品应用

功能&接口



- CPU x2
 - 额定运行频率400MHz, 各自带独立FPU(单精度浮点)
- KPU x1: 神经网络加速器
 - 卷积、池化、激活
- APU x1: 语音处理单元
 - Beam-forming、降采样、FIR、FFT
 - 最多支持8路音频数据, 16个方向
- SRAM x1: 8MB
 - 其中2MB为KPU专用,剩余6MB由用户程序和模型数据共用
- DVP x1: 摄像头接口(8位并口)
 - 输入: RGB565、YUV422
 - 输出: RGB565、RGB888、YUV422的Y分量
 - 支持的最大分辨率1920*1023
- DMA x6: 6个通道

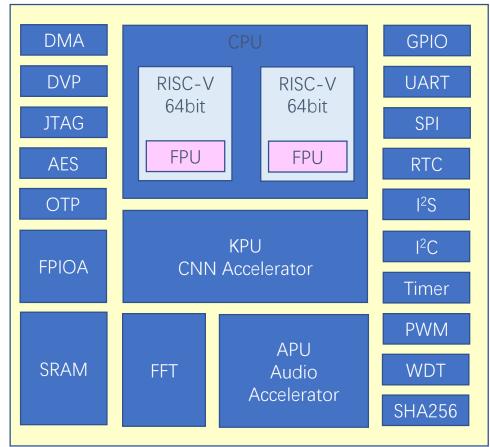


功能&接口



- FPIOA x1: 现场可编程IO阵列
 - 用于将芯片内部功能映射到48个物理IO上
- GPIO x40
- UART x4
 - 推荐使用UART1、UART2、UART3, 波特率达5Mbps
 - 推荐UARTO仅用于调试
- SPI x4
 - SPIO、SPI1、SPI3为master, SPI2为slave
 - SPI3固定接片外FLASH, SPI0通常接屏幕
 - Master时钟速度达80MHz, slave时钟速度达30MHz
- 12S x3
 - 每个I2S最多可接8路音频数据
 - I2SO可以连接到APU
- 12C x3
- 定时器 x3, WDT x2, RTC x1
- FFT x1, SHA256 x1, AES x1

TSMC 28nm

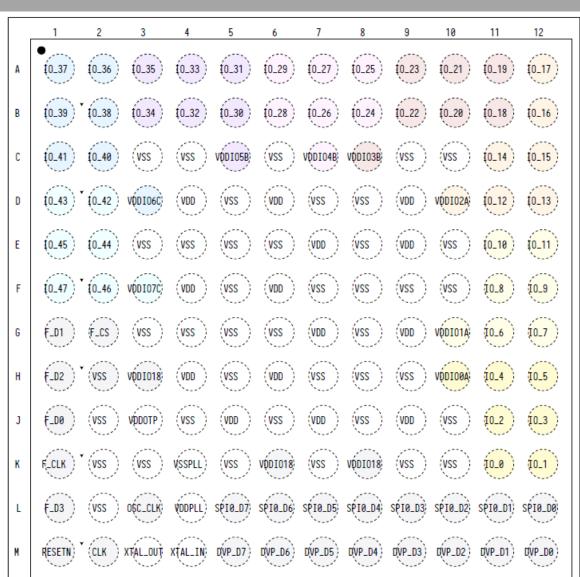


硬件设计



• 电源

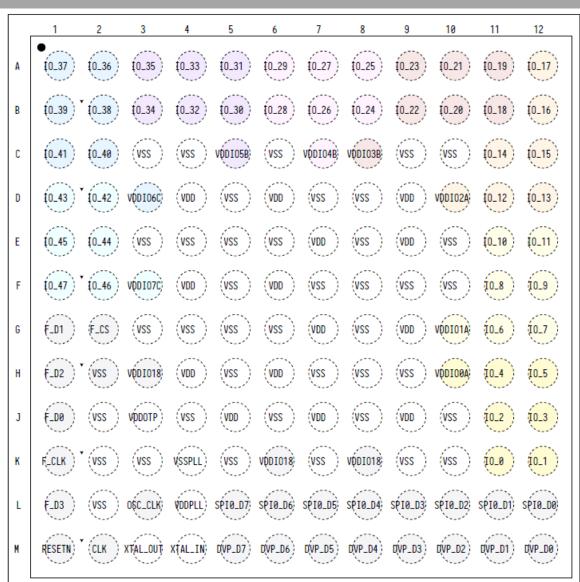
电源域组	电源域	支持电压 (V)	互联特性	电源名称
A	0	3.3或1.8	组内互联,组间独立	VDDI00A
Α	1	3.3 或 1.8	组内互联,组间独立	VDDI01A
Α	2	3.3 或 1.8	组内互联,组间独立	VDDI02A
В	3	3.3 或 1.8	组内互联,组间独立	VDDI03B
В	4	3.3 或 1.8	组内互联,组间独立	VDDI04B
В	5	3.3 或 1.8	组内互联,组间独立	VDDI05B
C	6	3.3 或 1.8	组内互联,组间独立	VDDI06C
C	7	3.3 或 1.8	组内互联,组间独立	VDDI07C
低压 IO	低压 I0	1.8	无特殊要求	VDDI018
OTP	OTP	1.8	无特殊要求	VDDOTP
PLL	PLL	0.9	无特殊要求	VDDPLL
数字核心	数字核心	0.9	无特殊要求	VDD



硬件设计

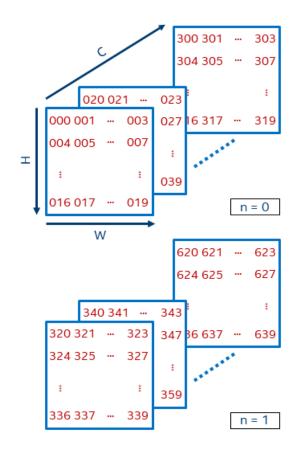


- FLASH
 - 专用SPI3, 电平1.8V
- 时钟
 - 注意要将L3和M2短接
- DVP与SPIO
 - 都有专用数据引脚
 - 二者存在绑定关系,要么都用专用引脚,要么都不用
- 特殊引脚
 - IO16用于boot模式选择,上电复位时,拉高进入 FLASH启动,拉低进入ISP模式
 - IO4和IO5默认为UARTO, 用于程序下载



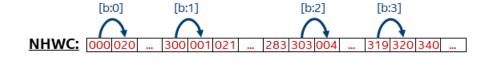


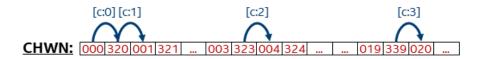
- 网络输入或者中间的feature map尺寸不大于320*256,不小于4*4
- 与tensorflow支持的数据排布格式NHWC不同, KPU支持的数据排布格式为NCHW



Physical data layout NCHW, NHWC, and CHWN layouts



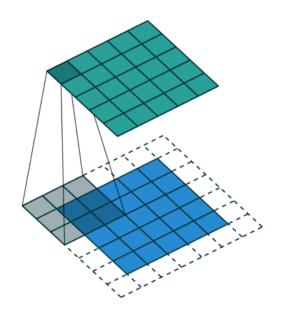






• 支持的普通二维卷积

ksize	stride	required tensorflow padding	
1*1	1	same或valid	
3*3	1	same	
3*3	2	卷积前将feature map手动padding一圈,然后卷积时padding方式用valid	





Input Volume (+pad 1) (7x7x3)	Filter WO (3x3x3)	Filter W1 (3x3x3)	Output Volume (3x3x2)
x[:,:,0]	w0[:,:,0]	w1[:,:,0]	o[:,:,0]
	1 1 -1	-1 -1 0	1 0 -3
0 0 1 1 2 2 0	-1 0 1	-1 1 0	-6 1 1
0 0 1 1 0 0 0	-1 -1 0	-1 1 0	4 -3 1
0 1 1 0 1 0 0	w0[:,:,1]	w1[:,:,1]	0[:,:,1]
0 1 0 1 1 1 0	-1 0 -1	1 -1 0	-1 -6 -4
0 0 2 0 1 0 0	0 0 -1	-1 0 -1	-2 -3 -4
000000	1 -1 0	-1 0 0	-1 -3 -3
x[:,:,1]	w0[:,4,2]	w1[:,:,2]	
0000000	0 1 0	-1 0 1	
0 1 1 1 2 0 0	1 8 1	1 0 1	
0 8 2 1 1 2 8	0 -1 1	0 -1 0	
0 1 2 0 0 2 0		14 (4 4 4)	
0 0 2 1 2 1 0	Bias b0 (1x1x1) b0(:,:,0]	Bias b1 (1x1x1)	
0 2 0 1 2 0 0	1	b1[:,:,0] 0	
0 0 0 0 0 0			
x(:,:,2] 0 0 0 0 0 0		toggle m	ovement
0 2 0 2 0 2 0			
0 0 0 1 2 1 0			
0 1 0 2 2 1 0			
0 2 0 2 0 0 0			
0 0 0 1 1 2 0			
0 0 0 0 0 0 0			



• 支持的depthwise二维卷积

ksize	stride	required tensorflow padding	
3*3	1	same	
3*3	2	卷积前将feature map手动padding一圈,然后卷积时padding方式用valid	

• 支持的pooling

ksize	stride	required tensorflow padding
2*2	1	same
2*2	2	如果feature map size是ksize的整数倍,same或valid;否则valid
4*4	4	如果feature map size是ksize的整数倍,same或valid;否则valid

模型转换



用于将主流训练框架按照特定限制规则训练出来的浮点模型量化转化为K210可运行的格式

- 支持的训练框架
 - Tensorflow、Caffe、Paddlepaddle
- 量化位数
 - 8bit、16bit
- 支持的操作
 - Conv2d、DepthwiseConv2d、FullyConnected、Add、MaxPool2d、AveragePool2d、GlobalAveragePool2d、BatchNormalization、BiasAdd、Relu、Relu6、LeakyRelu、Concatenation、L2Normalization、Softmax、Flatten、ResizeNearestNeighbor