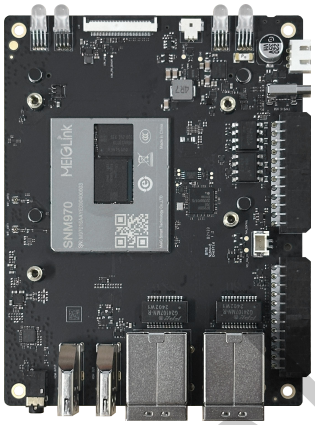


MeiG Pi-QCS8550

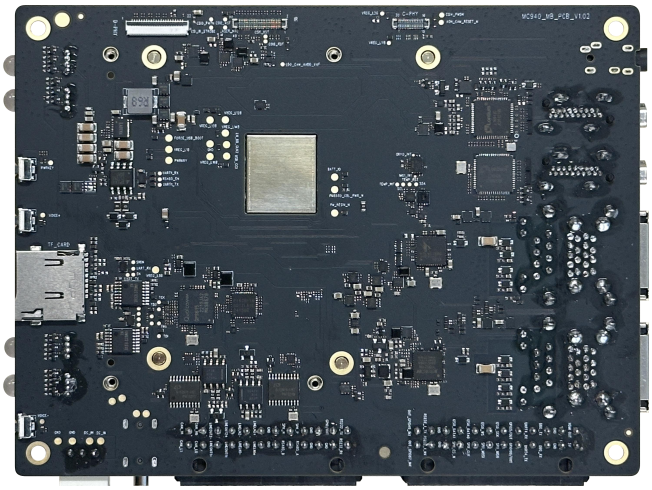
AI 高算力模组

面向开发者套件

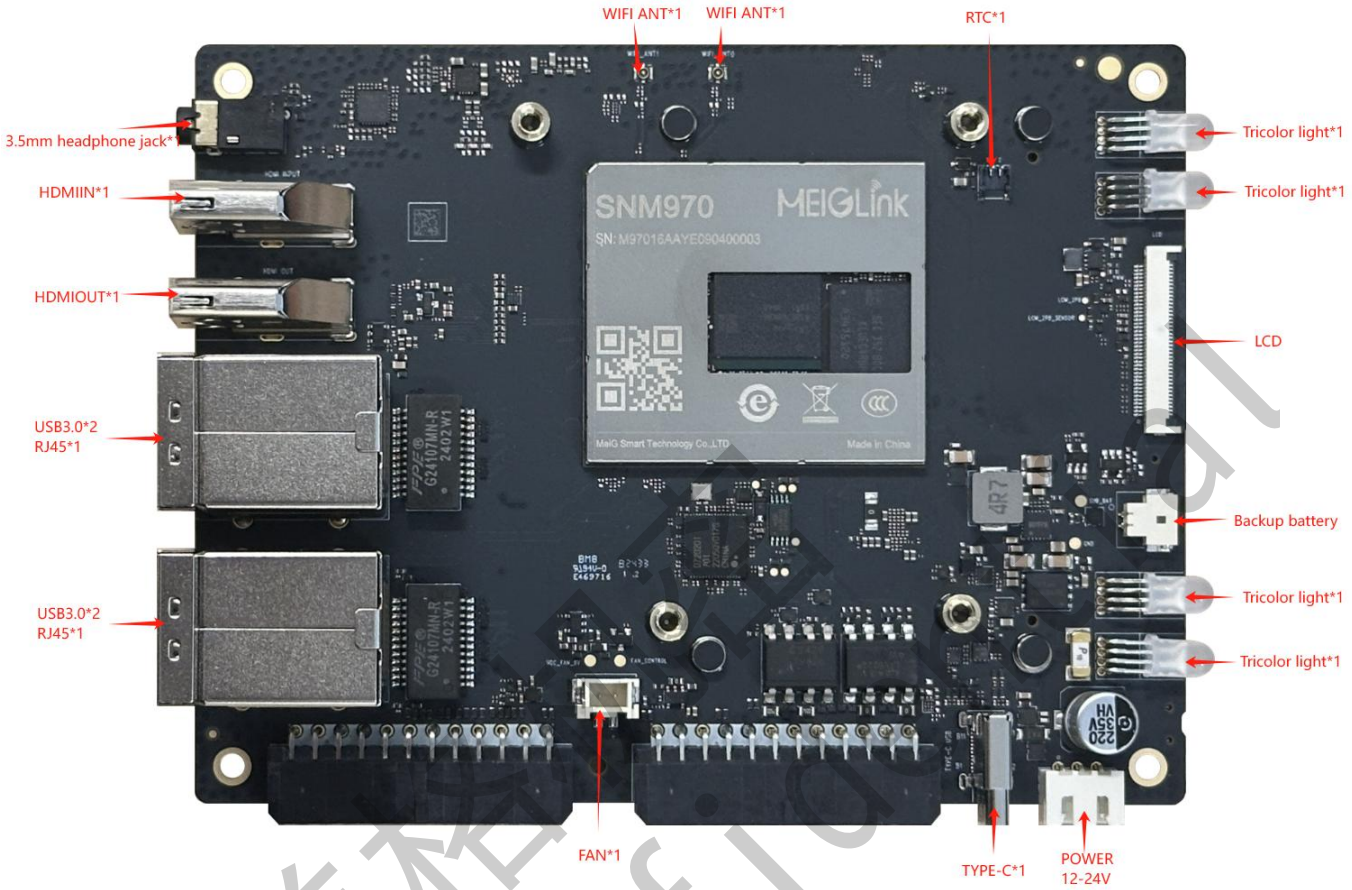
MeiG Pi 开发套件是一款基于美格 SNM970(QCS8550)开发设计的高算力开发套件，采用 4nm 制程，集成了 Android 系统。QCS8550 算力高达 48 Tops(int8)，可兼容 SNM932(QCS6490) & SNM950(QCS8250) & SNM960(SM8475)，拥有强大的图形处理能力和视频编解码能力，以及丰富的外围接口。

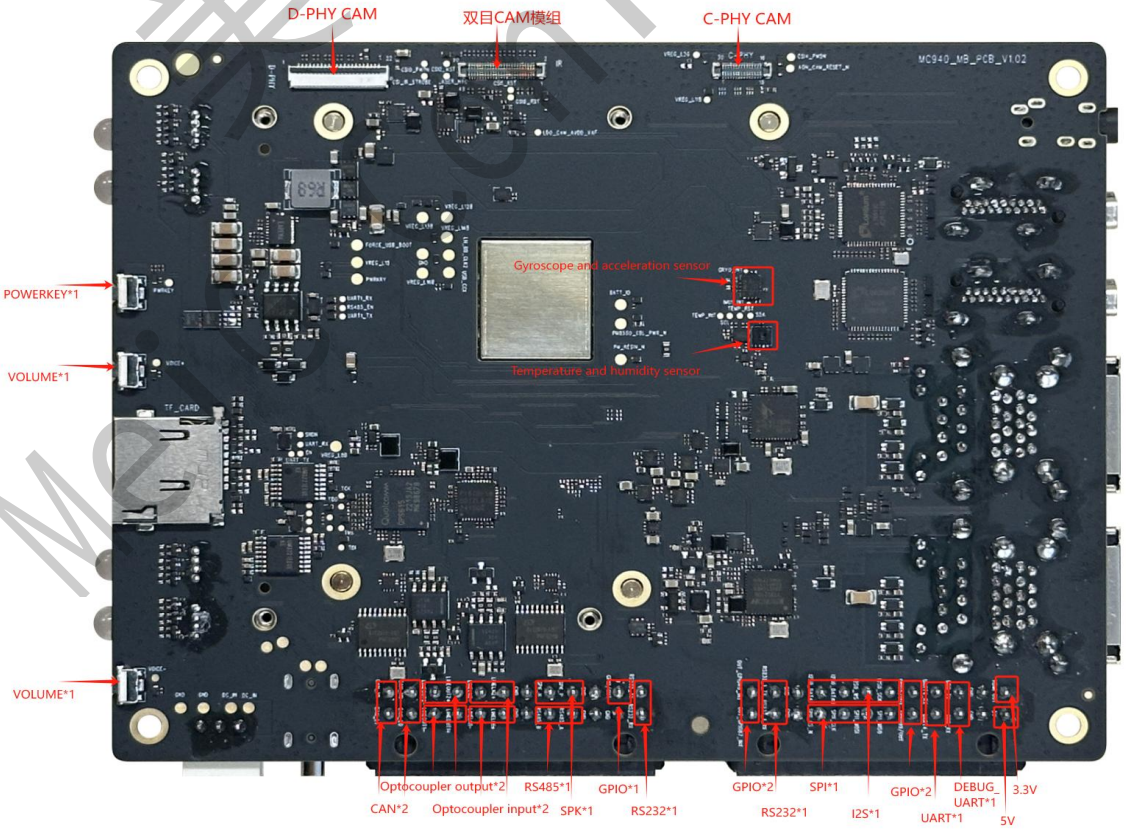
开发套件可面向用于工业相机、智算服务器、边缘智算盒、工控机、无人机、机器人、视频记录仪、AIOT 等多种类型产品。

MeiG Pi 开发套件概览图



MeiG Pi 开发套件概览图





技术参数

平台	QCS8550
AI 性能	~48 Tops（int8）
CPU	1x Kryo Prime 3.2 GHz + 4x Kryo Gold 2.8 GHz + 3x Kryo Silver 2.0 GHz
GPU	Adreno™ 740 @ 680MHz
OS	Android 13
运行内存	256GB UFS4.0 + 16GB LPDDR5x （默认） 256GB UFS4.0 + 24GB LPDDR5x （可选） 512GB UFS4.0 + 24GB LPDDR5x （可选）
视频编解码	Decode : 4K@240fps or 8K@60fps H.264/H.265/VP9 Encode : 4K@120fps or 8K@30fps H.264/H.265

硬件接口

C-PHY-Camera 模组	x 1，IMX586(48M)
D-PHY-Camera 模组	x 1，IMX577(12M)
双目结构光 Camera 模组	x 1，IMX577(12M)+2*OV9282(2M)
LCD	x 1，MIPI 接口，默认 ST7703，720*1280（客户可以自己定义转接 FPC，分辨率可提高）
USB Type-A	x 4，USB3.0，host 模式 （4 路总带宽可以支持到 5Gbps）
USB Type-C	x 1，USB3.0，支持 DP1.4
HDMI OUT	x 1，支持 HDMI 1.4，支持 1080P@60Hz
HDMI IN	x 1，支持 HDMI 1.4，支持 4K@30Hz
LAN(RJ45)	x 2，一路 2.5Gbps，一路 1Gbps
CAN	x 2
RS232	x 2，传输串口信息
RS485	x 1，传输串口信息
光耦输入	x 2，5~30V
光耦输出	x 2，5~30V
TF 卡	x 1
I2S	x 1，3.3V
SPI	x 1，3.3V
GPIO	x 5，3.3V
UART	x 1，3.3V
调试 UART	x 1，用于 debug
耳机	x 1，3.5mm
SPK	x 1，1.65W
三色灯	x 4，三色指示灯：红、绿、蓝
KEY	x 3，power/vol+/vol-
电源接口 12V/24V	x 1，外部适配器供电
RTC	x 1，CR2032(225mAh)
ANT	x 2 WIFI/BT 天线
工作温度	-30℃~+75°C
存储温度	-40°C ~ +90°C
尺寸	100 x 130 x 39.25mm

LLM 模型测试数据

模型格式：qnn，使用 genie-t2t-run，CTX=1024，256GB UFS4.0 + 24GB LPDDR5x。

启动时间 **（定义为 SDK 完成整个初始化耗时，主要组成部分为模型加载与内存申请）**					
模型	Llama2-7B-Chat (CTX=1024)	Qwen1.5-1.8B-Chat (CTX=1024)	Qwen1.5-0.5B-Chat (CTX=1024)	Qwen2.5-3B(context=1024)	qwen1.5-7B (context=1024)
耗时	1760ms	954ms	800ms	1257ms	
处理速度 **（分为编码速度 Prefill 和解码速度 Decode）**					
模型	Llama2-7B-Chat (CTX=1024)	Qwen1.5-1.8B-Chat (CTX=1024)	Qwen1.5-0.5B-Chat (CTX=1024)	Qwen2.5-3B(context=1024)	qwen1.5-7B (context=1024)
Prefill (token/s)	510	1969	4276	1517	
Decode (token/s)	10~11	32	84	25	
内存占用 **（进程当前正在使用的物理内存量，不包括交换空间，即 RSS）**					
模型	Llama2-7B-Chat (CTX=1024)	Qwen1.5-1.8B-Chat (CTX=1024)	Qwen1.5-0.5B-Chat (CTX=1024)	Qwen2.5-3B(context=1024)	qwen1.5-7B (context=1024)
占用	3.6G	1.6	1.2G	0.9G	
CPU 负载 **（推理过程中，在单元时间内（100 毫秒），进程实际工作所占的百分比）**					
模型	Llama2-7B-Chat (CTX=1024)	Qwen1.5-1.8B-Chat (CTX=1024)	Qwen1.5-0.5B-Chat (CTX=1024)	Qwen2.5-3B(context=1024)	qwen1.5-7B (context=1024)
峰值	35% / 100%	34% / 100%	15% / 100%	34% / 100%	
推理	30% / 100%	30% / 100%	12% / 100%	31% / 100%	
NPU 负载 **（在一定时间内，Genie 或 AidGen 进程实际使用 NPU 时间所占百分比。）**					
模型	Llama2-7B-Chat (CTX=1024)	Qwen1.5-1.8B-Chat (CTX=1024)	Qwen1.5-0.5B-Chat (CTX=1024)	Qwen2.5-3B(context=1024)	qwen1.5-7B (context=1024)
负载	99%	97%	96%	95%	

数据仅供参考

YOLO 测试数据

模型格式：dlc
cpu gpu：float32
npu：int8
输入图片尺寸：640*640
输入组数：1000 组，256GB UFS4.0 + 16GB LPDDR5x

启动时间(Create Network(s))						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
耗时	220ms	14709ms	26ms	133ms	5284ms	18ms
处理速度						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
Total Inference(inf/s)	0.29	24.39	138.89	0.57	40	129.87
Prefill (token/s)	/	/	/	/	/	/
Decode (token/s)	/	/	/	/	/	/
内存占用 （进程当前正在使用的物理内存量，不包括交换空间，即 RSS）						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
占用	222MB	135MB	42MB	127MB	107MB	35M
CPU 负载 （推理过程中，在单元时间内（1 秒），进程实际工作所占的百分比）						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
峰值	8%	8%	2%	13%	9%	3%
推理	5%	4%	2%	6%	8%	2%
GPU 负载						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
峰值	/	38%	/	/	27%	/
推理	/	35%	/	/	25%	/
NPU 负载 （在一定时间内，Genie 或 AidGen 进程实际使用 NPU 时间所占百分比。）						
模型	YOLOV5s			YOLOV8n		
backend	CPU	GPU	NPU	CPU	GPU	NPU
负载	/	/	40%	/	/	52%

数据仅供参考

Distibert sst2 测试数据

模型格式：dlc
cpu gpu：float32
npu：int8
输入文字长度：18 个单词
输入组数：500 组，256GB UFS4.0 + 16GB LPDDR5x

启动时间 （定义为 SDK 完成整个初始化耗时，主要组成部分为模型加载与内存申请）			
模型	distilbert		
backend	CPU	GPU	NPU
耗时	520.452ms	3307.305ms	31.017ms
处理速度			
模型	distilbert		
backend	CPU	GPU	NPU
Total Inference(inf/s)	7	5	34
Prefill (token/s)	/	/	/
Decode (token/s)	/	/	/
内存占用 （进程当前正在使用的物理内存量，不包括交换空间，即 RSS）			
模型	distilbert		
backend	CPU	GPU	NPU
占用	146MB	439MB	58MB->317MB（运行过程中不断增加）
CPU 负载 （推理过程中，在单元时间内（1 秒），进程实际工作所占的百分比）			
模型	distilbert		
backend	CPU	GPU	NPU
峰值	63%	12%	15%
推理	61%	11%	14%
GPU 负载			
模型	distilbert		
backend	CPU	GPU	NPU
峰值	/	94%	/
推理	/	94%	/
NPU 负载 （在一定时间内，Genie 或 AidGen 进程实际使用 NPU 时间所占百分比。）			
模型	distilbert		
backend	CPU	GPU	NPU
负载	/	/	84%

数据仅供参考

Whisper 测试数据

模型格式：tflite

录音时间：5 秒，256GB UFS4.0 + 16GB LPDDR5x

启动时间 **（定义为 SDK 完成整个初始化耗时，主要组成部分为模型加载与内存申请）**			
模型	whisper-tiny-en		
backend	CPU	GPU	NPU
耗时	17.06ms	/	/
音频特征提取（非神经网络）			
模型	log_mel_spectrogram		
backend	CPU	GPU	NPU
耗时	1200ms	/	/
处理速度			
模型	whisper-tiny-en		
backend	CPU	GPU	NPU
Total Inference(inf/s)	0.714	/	/
Prefill (token/s)	/	/	/
Decode (token/s)	/	/	/
内存占用 **（进程当前正在使用的物理内存量，不包括交换空间，即 RSS）**			
模型	whisper-tiny-en		
backend	CPU	GPU	NPU
占用	296.43MB	/	/
CPU 负载 **（推理过程中，在单元时间内（100 毫秒），进程实际工作所占的百分比）**			
模型	whisper-tiny-en		
backend	CPU	GPU	NPU
峰值	23%	/	/
推理	5%	/	/

数据仅供参考