



# **T527 NPU 常见网络性能 测试报告**

版本号: 2.0  
发布日期: 2024.11.13

## 版本历史

版本号	日期	制/修订人	内容描述
1.0	2023.04.10	AWA2090	NPU 常见网络性能测试报告初始版本
2.0	2024.11.13	AWA1382	补充常见模型的数据。



# 目 录

<b>1 前言</b>	<b>1</b>
1.1 文档简介 .....	1
1.2 目标读者 .....	1
1.3 适用范围 .....	1
1.4 术语，缩略语及概念 .....	1
<b>2 环境搭建</b>	<b>2</b>
<b>3 测试步骤</b>	<b>3</b>
<b>4 公开模型的性能测试</b>	<b>5</b>
<b>5 公开模型的资源情况</b>	<b>6</b>
<b>6 附录</b>	<b>7</b>



# 1 前言

## 1.1 文档简介

介绍 NPU 模块的常见网络性能的测试报告，为 NPU 驱动、算法开发提供参考。

## 1.2 目标读者

本文档（本指南）主要适用于以下人员：

- 技术支持工程师
- 软件开发工程师
- AI 应用案客户

## 1.3 适用范围

硬件平台：T527 平台

软件平台：Linux5.15 系统（buildroot）、Android 13 系统

## 1.4 术语，缩略语及概念

术语	解释说明
NPU	Neural Network Processing Unit，神经网络处理器，采用“数据驱动并行计算”的架构，擅长处理视频、图像类的海量多媒体数据。
VIPLite	VIPLite 版本 NPU 驱动。
Unified	Unified 版本 NPU 驱动。
带宽	发送信号中含有的有效成分的效率范围，可以用来标识信号传输的数据传输能力，简写 BW。

## 2 环境搭建

修改设备树配置 NPU 模块的时钟频率为 546MHz/696MHz 进行测试：

```
interrupt-names = "npu";  
- npu-vf = <696>;  
+ npu-vf = <546>;  
+ ##npu-vf = <696>;  
power-domains = <&pd A523_PD_NPU>;
```

编译 NPU 驱动并传到板端挂载（编译进内核不需要）：

```
##挂载VIPLite版本NPU驱动  
insmod vipcore.ko  
##挂载Unify版本NPU驱动  
insmod galcore.ko
```

挂载动态库，公版方案中已预置动态库，若无预置使用adb push方式将动态库推到板端/usr/lib或者其他目录下：

VIPLite 版本动态库有：

```
├── libVIPLite.so  
└── libVIPuser.so
```

Unified 版本动态库有：

```
├── libCLC.so  
├── libGAL.so  
├── libGLSLC.so  
├── libNNGPUBinary.so  
├── libNNVXCBinary.so  
├── libOpenVX.so  
├── libOpenVX.so.1  
├── libOpenVX.so.1.3.0  
├── libOpenVXU.so  
├── libOvx12VXCBinary.so  
├── libOvxGPUVXCBinary.so  
└── libovxlib.so
```

若推到其他目录下时，请配置环境变量（使用静态库编译例程时不需要）：

```
export LD_LIBRARY_PATH=/xxx/xxx
```

### 3 测试步骤

本文以 VIPLite 版本 NPU 驱动为例，使用 vpm\_run 程序进行测试。vpm\_run 程序使用方法请参考《NPU\_模型部署\_开发指南》。公版方案中已预置 vpm\_run 程序，若无预置参考《NPU\_模型部署\_开发指南》的《模型运行工具 vpm\_run 使用说明》一节进行编译。

1. 将测试用例以及模型和输入使用 adb push 传入板端；
2. 配置 sample.txt 指定模型与输入；

```
[network]
./network_binary.nb
[input]
./input.dat
```

3. 运行测试用例：

```
chmod +x vpm_run
##运行例程 -l 100: 推理100次
./vpm_run -s sample.txt -l 100
```

4. 输出结果，分析 LOG 得出结论。

```
/xx/test/vpm_run # ./vpm_run -s sample.txt -l 100
loop_count=100, device_id=0, file_name=sample.txt
test started.

init vip lite, driver version=0x00010d00...
VIPLite driver software version 1.13.0.0-AW-2023-10-19
vip lite init OK.

cid=0x10000016, device_count=1
device[0] core_count=1
init test resources, batch_count: 1 ...
create/prepare networks ...
batch i=0, binary name: ./network_binary.nb
input 0 dim 224 224 3 1, data_format=2, quant_format=2, name=input[0], scale=0.007843, zero_point=128
output 0 dim 2 1 0 0, data_format=2, name=uid_1_out_0, scale=0.065554, zero_point=128
nbg name=./network_binary.nb
create network 0: 1763 us.
memory pool size=1092608byte
network core count=1
input 0 name: ./input_0.dat
prepare network 0: 2214 us.
```

```
batch: 0, loop count: 1
start to run network=./network_binary.nb
run time for this network 0: 15075 us.
run network done...
#第1次推理时间
profile inference time=14678us, cycle=10127721
***** nb TOP5 *****
...
...
batch: 0, loop count: 100
start to run network=./network_binary.nb
run time for this network 0: 15125 us.
run network done...
#第100次推理时间
profile inference time=14701us, cycle=10101912
***** nb TOP5 *****
--- Top5 ---
...

#分析log, 可知平均推理时间为14678us
batch 0, profile avg inference time=14678us, cycle=10127721
destroy teset resource batch_count=1
```



## 4 公开模型的性能测试

测试各模型的帧率（推理 1000 次）、带宽占用等情况，对 546MHz 以及 696MHz 频率的 NPU 模块以及 924GHz 的 DRAM 进行测试。

模型名称	输入分辨率	量化精度	FPS(546)	FPS(696)	ReadBW(MB)	WriteBW(MB)
yolov3	3x416x416	uint8	13.02	15.74	110.83	36.82
yolov4-tiny	3x416x416	uint8	92.23	115.11	10.77	4.02
yolov5s-sim	3x640x640	uint8	16.74	20.56	47.71	23.40
yolov5m	3x608x608	uint8	9.16	11.14	112.14	56.69
yolov7	3x640x640	uint8	4.23	5.18	276.81	72.83
yolov9c	3x640x640	uint8	3.98	4.82	270.55	125.66
ssd_mobilenet_v1	300x300x3	uint8	125.03	155.35	9.54	2.59
alexnet	3x227x227	uint8	34.97	44.45	24.20	0.81
inception_v3	299x299x3	uint8	29.13	35.88	34.24	5.26
inception_v4	299x299x3	uint8	14.35	17.55	79.29	13.05
mobilenetv1_1.0	224x224x3	uint8	274.57	347.58	4.36	1.29
mobilenetv2-12	3x224x224	uint8	293.77	348.43	5.47	2.16
resnet50v2	3x224x224	uint8	30.96	38.49	33.26	8.80
fairmot_crowdhuman	3x608x1088	uint8	2.57	3.04	361.58	459.55
yolact_resnet50	3x550x550	uint8	4.09	4.91	263.25	203.23
openpose	3x184x184	uint8	15.53	19.72	61.88	4.34
alphaPose_hard68	3x256x192	uint8	4.69	5.94	55.23	13.68
handpose_x	3x256x256	uint8	78.08	94.54	18.34	9.84
PFLD98-sim	3x112x112	uint8	610.87	778.21	1.18	0.06
Swin_MLP	3x224x224	uint8	7.09	8.7	74.29	36.59
Swin_Transformer	3x224x224	uint8	3.33	4.01	253.84	63.22
Swin_Transformer_v2	3x224x224	uint8	2.29	2.78	357.66	71.91
SimpleViT-sim	3x256x256	uint8	14.43	17.98	62.09	6.00
T2TViT-sim	3x256x256	uint8	1.31	1.6	454.52	220.42



## 5 公开模型的资源情况

以下为测试的公开模型的资源占用情况，内存占用统计使用了 VmHWM 值：

模型名称	输入分辨率	量化精度	模型大小 (MB)	内存占用 (MB)
yolov3	3x416x416	uint8	37.43	38.08
yolov4-tiny	3x416x416	uint8	3.48	3.69
yolov5s-sim	3x640x640	uint8	6.48	7.63
yolov5m	3x608x608	uint8	14.26	15.11
yolov7	3x640x640	uint8	24.28	26.67
yolov9c	3x640x640	uint8	18.27	20.46
ssd_mobilenet_v1	300x300x3	uint8	4.49	4.94
alexnet	3x227x227	uint8	45.49	45.66
inception_v3	299x299x3	uint8	16.61	16.89
inception_v4	299x299x3	uint8	28.85	29.18
mobilenetv1_1.0	224x224x3	uint8	2.96	3.14
mobilenetv2-12	3x224x224	uint8	2.96	3.20
resnet50v2	3x224x224	uint8	16.52	16.71
fairmot_crowdhuman	3x608x1088	uint8	15.89	17.41
yolact_resnet50	3x550x550	uint8	24.13	24.89
openpose	3x184x184	uint8	31.76	31.95
alphaPose_hard68	3x256x192	uint8	2.73	20.67
handpose_x	3x256x256	uint8	2.69	2.93
PFLD98-sim	3x112x112	uint8	1.11	1.27
Swin_MLP	3x224x224	uint8	20.73	22.16
Swin_Transformer	3x224x224	uint8	38.89	42.66
Swin_Transformer_v2	3x224x224	uint8	47.22	52.07
SimpleViT-sim	3x256x256	uint8	56.98	57.50
T2TViT-sim	3x256x256	uint8	70.66	72.05

## 6 附录

公开模型获取途径如下：

分类	模型	获取途径
目标检测	yolov3	<a href="#">cfg、weights</a>
目标检测	yolov5s-sim	<a href="#">下载</a>
目标检测	yolov9c	<a href="#">下载</a>
目标检测	ssd_mobilenet_v1	<a href="#">下载</a>
分类	alexnet	<a href="#">下载，权重</a>
分类	resnet50v2	<a href="#">下载</a>
分类	mobilenetv1_1.0	<a href="#">下载</a>
分类	mobilenetv2-12	<a href="#">下载</a>
分类	inception_v3	<a href="#">下载</a>
分类	inception_v4	<a href="#">下载</a>
多目标跟踪	fairmot_crowdhuman	<a href="#">下载</a>
实例分割	yolact_resnet50	<a href="#">下载</a>
人体姿态估计	openpose	<a href="#">下载</a>
人体姿态估计	alphaPose_hard68	<a href="#">下载</a>
手部关键点	handpose_x	<a href="#">下载</a>
人脸关键点	PFLD98-sim	<a href="#">下载</a>
Transformer 骨干	Swin_Transformer	<a href="#">下载</a>
Transformer 骨干	Swin_Transformer_v2	<a href="#">下载</a>
MLP	Swin_MLP	<a href="#">下载</a>
Transformer 骨干	T2TViT-sim	<a href="#">下载</a>
Transformer 骨干	SimpleViT-sim	<a href="#">下载</a>




## 著作权声明

版权所有 © 2024 珠海全志科技股份有限公司。保留一切权利。

本文档及内容受著作权法保护，其著作权由珠海全志科技股份有限公司（“全志”）拥有并保留一切权利。

本文档是全志的原创作品和版权财产，未经全志书面许可，任何单位和个人不得擅自摘抄、复制、修改、发表或传播本文档内容的部分或全部，且不得以任何形式传播。

## 商标声明

、、**全志科技**、（不完全列举）均为珠海全志科技股份有限公司的商标或者注册商标。在本文档描述的产品中出现的其它商标，产品名称，和服务名称，均由其各自所有人拥有。

## 免责声明

您购买的产品、服务或特性应受您与珠海全志科技股份有限公司（“全志”）之间签署的商业合同和条款的约束。本文档中描述的全部或部分产品、服务或特性可能不在您所购买或使用的范围内。使用前请认真阅读合同条款和相关说明，并严格遵循本文档的使用说明。您将自行承担任何不当使用行为（包括但不限于如超压，超频，超温使用）造成的不利后果，全志概不负责。

本文档作为使用指导仅供参考。由于产品版本升级或其他原因，本文档内容有可能修改，如有变更，恕不另行通知。全志尽全力在本文档中提供准确的信息，但并不确保内容完全没有错误，因使用本文档而发生损害（包括但不限于间接的、偶然的、特殊的损失）或发生侵犯第三方权利事件，全志概不负责。本文档中的所有陈述、信息和建议并不构成任何明示或暗示的保证或承诺。

本文档未以明示或暗示或其他方式授予全志的任何专利或知识产权。在您实施方案或使用产品的过程中，可能需要获得第三方的权利许可。请您自行向第三方权利人获取相关的许可。全志不承担也不代为支付任何关于获取第三方许可的许可费或版税（专利税）。全志不对您所使用的第三方许可技术做出任何保证、赔偿或承担其他义务。