T-HEAD

TH1520 NPU
User Manual

Revision	1.0.0
Security	Secret
Date	2023-08-26

Trademarks and Permissions

The T-HEAD Logo and all other trademarks indicated as such herein are trademarks of T-HEAD (Shanghai) Semiconductor Co., Ltd. All other products or service names are the property of their respective owners.

Notice

The purchased products, services and features are stipulated by the contract made between T-HEAD and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

平头哥（上海）半导体技术有限公司 T-HEAD (Shanghai) Semiconductor Co., LTD

Address: 5th Floor Number 2 Chuan He Road 55, Number 366 Shang Ke Road, Shanghai free trade area, China
Website: www.t-head.cn

商标申明

平头哥的 LOGO 和其它所有商标归平头哥（上海）半导体技术有限公司及其关联公司所有，未经平头哥（上海）半导体技术有限公司的书面同意，任何法律实体不得使用平头哥的商标或者商业标识。

注意

您购买的产品、服务或特性等应受平头哥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的购买或使用范围之内。除非合同另有约定，平头哥对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。平头哥（上海）半导体技术有限公司不对任何第三方使用本文档产生的损失承担任何法律责任。

平头哥（上海）半导体技术有限公司 T-HEAD (Shanghai) Semiconductor Co., LTD

地址： 中国（上海）自由贸易试验区上科路 366 号、川和路 55 弄 2 号 5 层
网址： www.t-head.cn

Revisions

Rev	Description	Author(s)	Date
V1.0.0	Initial version	T-Head	2023-08-26

Contents

Figures & Tables

List of Abbreviations

Abbreviations	Full Spelling	Chinese Explanation
NPU	Neural-network Processing Unit	神经网络处理单元

1 Overview

NPU is a hardware based neural network accelerator which achieves high performance with low power. NPU is a key component for SoC targeting neural networks inference acceleration with support for variable precisions for data and weight. NPU supports weight compression and flexible low precision which allow neural networks to be run quickly. Figure & Table 1-1 shows the high level block diagram of NPU.

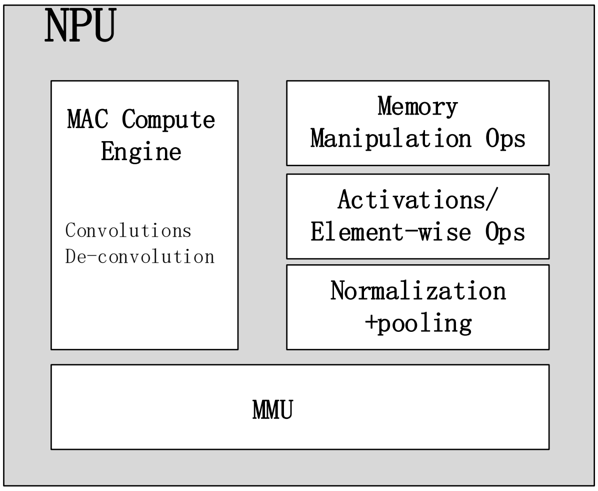

Figure & Table 1-1 NPU block function diagram

2 Main Features

NPU key features are:

- Acceleration of the most common layers of neural networks, listed in Figure & Table 2-1
- Low bandwidth operation
  - Support for a large range of low precision data formats
  - Grouping layers together to reduce memory bandwidth
  - Lossless weight data compression
- DRM security
- Interoperability
  - Support for a variety of memory formats designed to share data with a CPU, GPU or other modules

Figure & Table 2-1 Supported layers

Layer	Supported Via
Convolution	
Normal Convolution	Hardware
Dilated/Atrous Convolution	Hardware
Grouped Convolution	Hardware
Depthwise Convolution	Hardware
Convolution Transpose (Deconvolution)	Hardware
Fully Connected	
Fully Connected	Hardware
Normalization	
Batch normalization	Hardware
Local Response Normalization	Hardware
L2 Normalization	Software
Activation	
ReLU	Hardware
ReLU1	Hardware
ReLU6	Hardware
PReLU	Hardware
Clamped ReLU	Hardware

Leaky ReLU	Hardware
Tanh	Hardware
Sigmoid	Hardware
Logistic	Hardware
Pooling	
Max pooling	Hardware
Mean pooling	Hardware
Min pooling	Hardware
Elementwise Operation	
Negate	Hardware
Add/Subtract	Hardware
Multiply	Hardware
Max/min	Hardware
Memory Operation	
Permute	Hardware
Transpose	Hardware
Reshape	Hardware
Squeeze	Hardware
Flatten	Hardware
Space to Batch	Hardware
Batch to Space	Hardware
Depth to Space	Hardware
Space to Depth	Hardware
Spatial Resize Operation	
Pad	Hardware
Crop	Hardware
Bilinear resize	Hardware
Nearest neighbor resize	Hardware
Pre-processing	
Mean subtraction	Hardware

Post-processing	
Softmax	Software

3 Function Description

3.1 NPU Processing Order

Figure & Table 3-1 shows the order that layers are processed by NPU. The different layers that can be combined together in single pass through the hardware is called a "layer group". If the order of processing in the target network does not match the order shown in Figure & Table 3-1, then the operations will be split into different layer groups. For example, to perform local response normalization after pooling, the first layer group would perform the pooling layer, and the second would contain the local response.

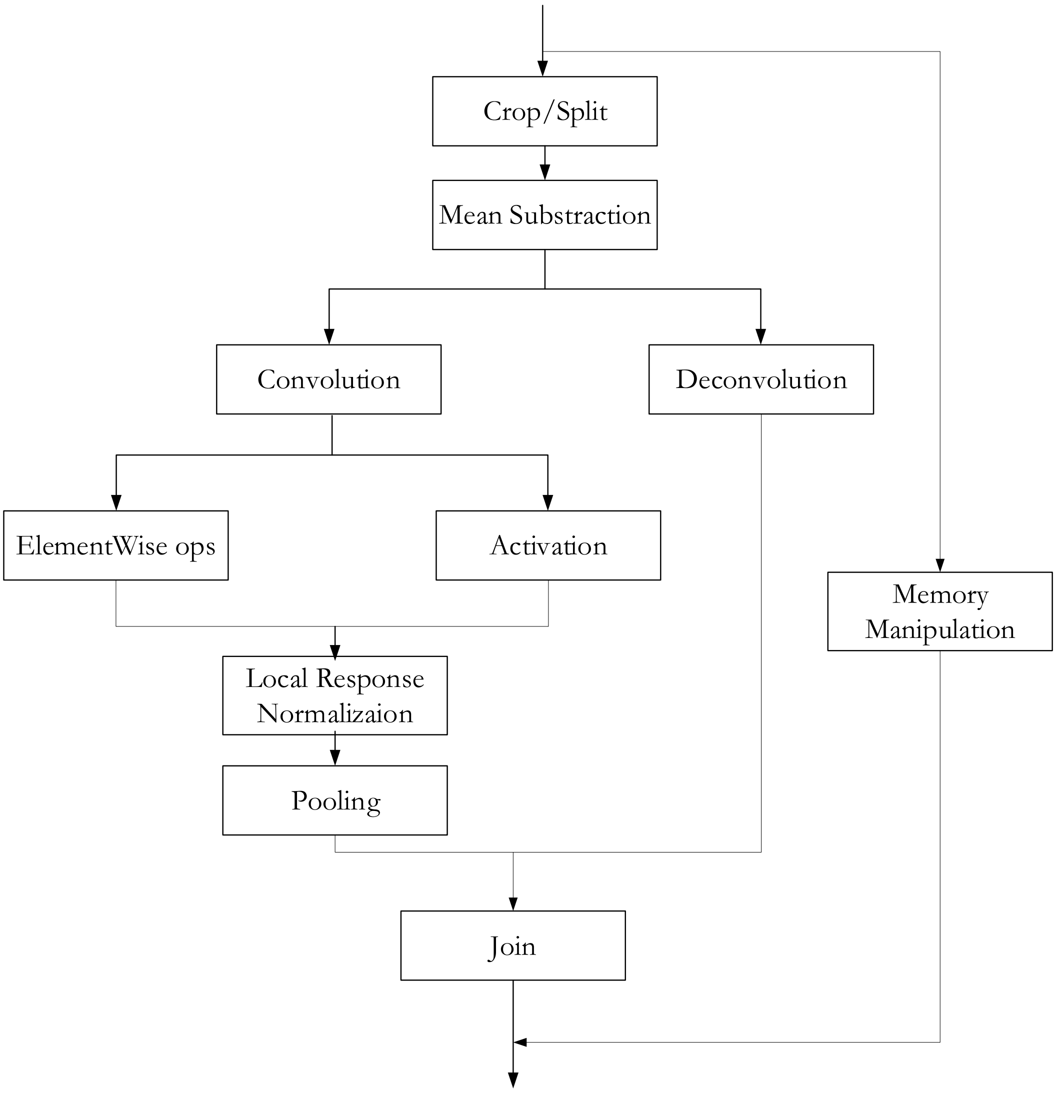

Figure & Table 3-1 NPU flow diagram

4 Usage

The power-up procedure of the core is as follows:

1. De-assert NPU resetn.
2. Power on NNA.
3. Enable NNA clock.
4. Wait at least 16 cycles.
5. Assert NPU resetn.
6. Wait at least 8 cycles.
7. Kick off NPU.