Technical **NOTE**

「DeePhi DPU でビデオ構造解析」
http://go.aps-web.jp/18-xilinx
QRコードで最新情報をご覧いただけます。

包括的AI推論アクセラレーションを提供
AI/MLモデルを高速化するハードウェア

ニューラルネットワークを開発する方々はCaffeやTensorFlowというフレームワークを使います。学習済みのネットワークをエッジ側に実装すると高スループット、低レイテンシ、優れたエネルギー効率を持つFPGA がよく採用されます。しかし、フレームワークの成果物となるネットワークモデルとその重み（係数）はどのようにFPGAに実装するかは課題でした。

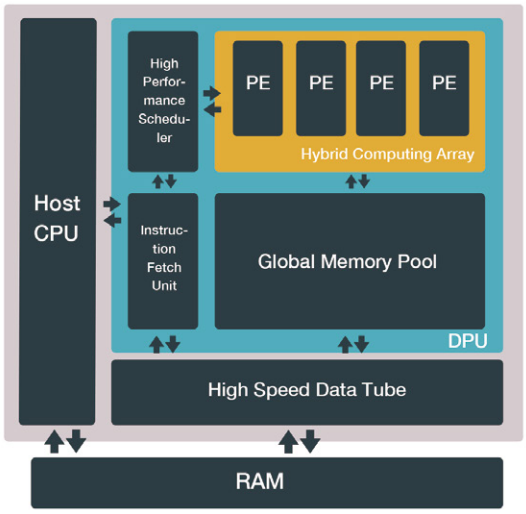
ザイリンクスはDeep Learning Processor Unit（DPU）と呼ばれる IP とフレームワークからZynq® UltraScale+™ MPSoC デバイスへ実装する開発ツールを提供しています。DPUはスケーラブルで、必要なスループットとユーザーロジックの規模に合わせて、そのIPを設定することができます。

表1に出ているアーキテクチャ記号に動作クロック周波数を掛けると最大計算力が得られます。例えば、B4096のIPコアを300MHzで動かすと、4096 * 300MHz = 1.2288TOPSになります。複数コアを実装することもできます。

Deep Neural Network Design Kit（DNNDK）のコンポーネンツは図2に示します。Pruningは計算量を減らせるオプショナルのツールです。現在のDPUはINT8で計算しているため、フレームワークが生成した重みと活性化関数の結果を8ビットに量子化するQuantizationツールがあります。Compilerはネットワーク構成をパーシング、データフローの最適化を行います。Assemblerは DPU IPに合わせた命令を生成します。Core APIでDPUに画像を入力、DPU実行、途中結果の取得することができます。ユーザーから見ると、Linuxデバイスをアクセスしているような感覚になります。DPUがサポートしている処理とDNNDKによってカスタム・ネットワークはFPGAに実装することが可能になります。

Arch	LUTs	FF	BRAM	DSP
B512	17951	28280	69.5	97
B800	20617	35065	87	141
B1024	22327	39000	101.5	193
B1152	22796	40276	117.5	193
B1600	26270	50005	123	281
B2304	29592	57549	161.5	385
B3136	33266	69110	203.5	505
B4096	37495	84157	249.5	641

表1：選択できるDPUアーキテクチャと使用するハードウェアリソース。



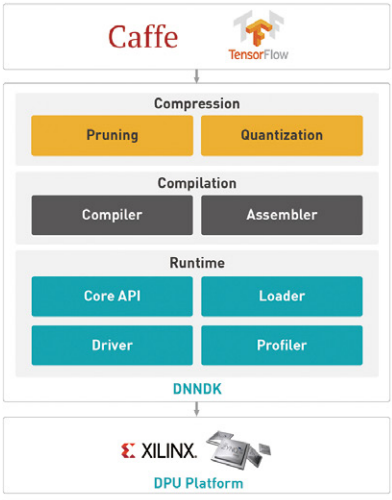


図1：Deep Learning Processor Unit（DPU）のブロック図 / 図2：DNNDK（マッピング開発ツール）

製品のお問い合わせは下記の販売代理店へどうぞ
■アヴネット株式会社　https://reach2.avnet.com/inquiry-japan-xpm.html
■株式会社PALTEK　https://www.paltek.co.jp/form/002.htm

XILINX

ザイリンクス株式会社
https://japan.xilinx.com/ml