この記事のURL：
http://techon.nikkeibp.co.jp/atcl/mag/15/320404/120400065/

日経テクノロジーonline

日経Automotive 2018年1月号

Features

NVIDIAの牙城に挑む、AI半導体

木村 雅秀　2017/12/11 00:00

出典：日経Automotive、2018年1月号、pp.50-55（記事は執筆時の情報に基づいており、現在では異なる場合があります）

人工知能（AI）を使った自動運転向けの半導体では、米NVIDIA社のGPUの存在感が大きい。ただ、GPUは高性能な半面、消費電力やコストに課題があり、大衆車に搭載するのは難しいとされる。このため、低電力・低コストの半導体を新たに開発する動きが活発化してきた。採用する自動車メーカーにとっては追い風といえる。新型の半導体はNVIDIA社の牙城を崩せるか。各社の取り組みを追った。

写真提供：アフロ、NVIDIA社

AIを使った自動運転の分野でGPUとそれ以外の半導体の競争が激しくなってきた（図1）。現在、AI向けの半導体では、米NVIDIA社のGPUの存在感が大きい。自動運転に欠かせないAIの学習用サーバーでは同社のGPUが標準的に使われているほか、AIの推論を担う車載半導体でもGPUが主役といえる。自動運転で先行する米Tesla社やドイツAudi社の車両もNVIDIA社のGPUを搭載している。

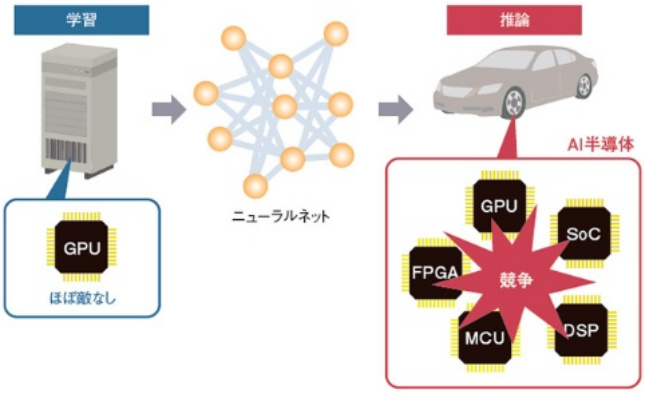

図1 AI半導体の競争が激化

AIの学習にはGPUが使われることが多いのに対し、車に搭載する推論用のチップはGPUのほか、さまざまな半導体が検討されている。消費電力やコストを巡る競争が激化しそうだ。本誌が作成。

GPUが選ばれる理由は、その性能の高さにある。「自動運転に必要な車載半導体の演算性能は120TOPS（Tera Operations Per Second）。現在のパソコン用CPUの2300倍以上だ」―。トヨタ自動車常務役員未来創生センター統括先進技術開発カンパニーExecutive Vice Presidentの奥地弘章氏はこう指摘する。120TOPSもの性能を持つ車載半導体はまだ存在しないが、GPUを搭載したNVIDIA社の車載AIコンピューター「DRIVE PX 2」は20TOPSと、現時点で入手可能なハードウエアの中では最も性能が高い。

その一方で、GPUは消費電力やコストが高く、開発用途や高級車向けには使えても、普及価格帯の量産車には使いにくいとの指摘が多い。DRIVE PX 2の消費電力は80～250Wと高く、「価格は数百万円もする」（ある半導体商社）という。NVIDIA社は低電力化を図った次世代SoC（System on Chip）「Xavier」を搭載した30TOPSの車載AIコンピューター「DRIVE PX Xavier」を、量産車向けに提供する予定だ。それでも消費電力は30Wと大きい。

こうした状況の中、多くの半導体メーカーはXavierの30TOPS、30Wという性能を一つの目安とし、それよりも低電力で低コストの車載半導体の開発を加速している。代表的な手法は「アクセラレーター」と呼ぶ専用回路を使ってGPUの機能を置き換える試みである。GPUが汎用的な演算器を多数並べているのに対し、アクセラレーターは特定の演算を高速・低電力に実行する専用回路を使う点が異なる。

打倒「Xavier」へ

ルネサスエレクトロニクスは2017年10月、トヨタとデンソーが2020年の実用化に向けて開発中の高速道路の自動運転技術「Highway Teammate」に、アクセラレーターを搭載した同社の車載SoC「R-Car H3」が採用されたと発表した（図2）。採用の決め手になったのは、「量産車に求められる性能と消費電力のバランス」（同社執行役員常務の大村隆司氏）だったという。

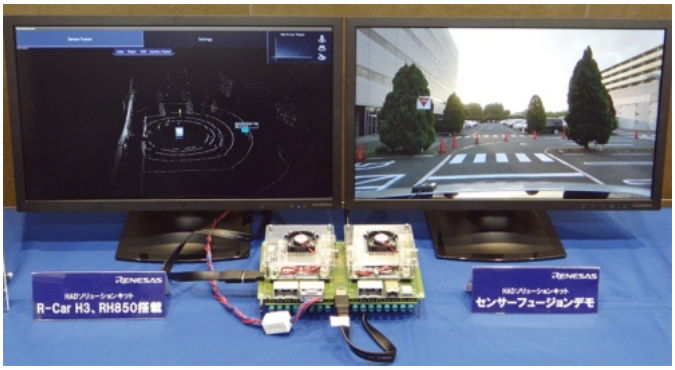

図2 トヨタが自動運転にルネサスのチップを採用

トヨタとデンソーが2020年の実用化に向けて開発中の自動運転技術「Highway Teammate」に、ルネサスの車載SoC「R-Car H3」と制御マイコン「RH850」が採用された。写真はR-Car H3とRH850を使ったデモの様子。本誌が撮影。

トヨタは2017年5月に自動運転でNVIDIA社と提携し、Xavierを使って数年以内にAIを活用した自動運転システムの市場投入を目指すとしていた。その一方で、トヨタはXavierの消費電力の高さを課題と捉えており、ルネサスのSoCも検討していた。Highway Teammateは現状ではAIを使わない見通しのため、今回のR-Car H3はXavierの対抗馬とはいえない。ただ、トヨタが2020年代前半の実用化を目指す一般道の自動運転技術「Urban Teammate」はAIを活用する可能性が高い。ルネサスはそこでXavierと対抗できる次世代SoCの投入を計画している。次世代SoCはR-Car H3と同様、アクセラレーターによる低電力化を追求する。

ルネサスのアクセラレーターは、GPUに比べて消費電力を約1/3に減らせる（図3）。カメラを使った画像認識では、画像信号処理やフィルター処理、認識処理などをアクセラレーターで実行することで、「GPUで10Wかかる処理を4Wで実現できた」（同社セーフティ・ソリューション事業部グローバルADASセンター課長の大塚聡氏）。さらに、アクセラレーターを動作時以外にスリープ（休止）させることで3Wに低減できたという。

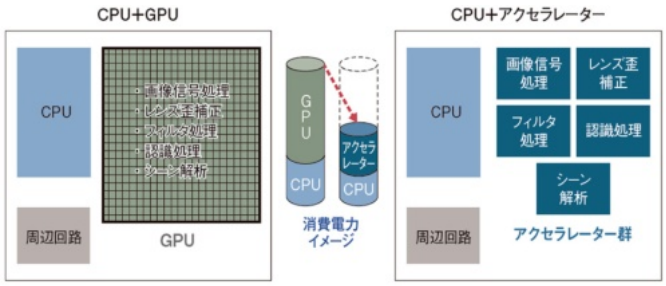

図3 アクセラレーターで消費電力を減らす

ルネサスはGPUの代わりにアクセラレーターを使って低消費電力化を図る。GPUで10Wかかる処理をアクセラレーターでは4Wで実現できる。さらにアクセラレーターの各ブロックを処理内容に応じてスリープ（休止）させることで、3Wまで低消費電力化できるという。ルネサスの資料を基に本誌が作成。

演算器の使用率を90%に

「AI向けの車載半導体で最も適しているのはアクセラレーターだ」―。AIを使った自動運転ソフトを開発するハンガリーAImotive社 Head of Japan OfficeのAxel Bialke氏はこう指摘する。同社はAIソフトを開発する立場から、GPUの消費電力の高さを問題視しており、より低電力な独自アクセラレーター「aiWare」の開発を進めている。

同社は菱洋エレクトロと技術協力し、物体認識用のニューラルネットをGPUとaiWareのそれぞれで動かすデモを2017年11月に見せた。アクセラレーターでもGPUと同様に物体を認識でき、歩行者を赤、車両を緑、自転車を黄色に分類できた（図4）。

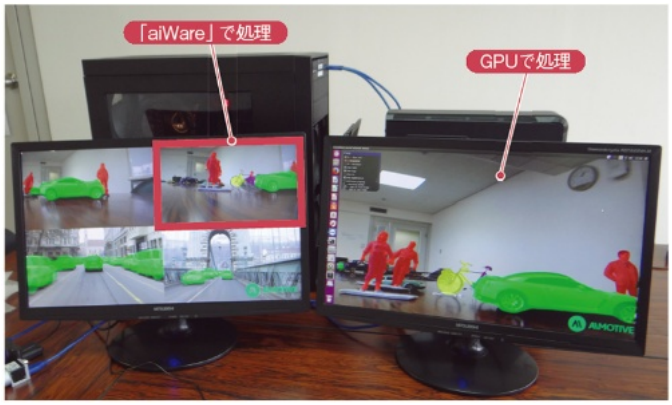

図4 同じニューラルネットをGPUとアクセラレーターで動かす

AImotive社はGPUとアクセラレーター「aiWare」で同じニューラルネットをそれぞれ動かすデモを見せた。写真の赤枠部分をaiWareで処理している。GPU

と同じように車両や歩行者を認識できるほか、消費電力をGPUの1/3～1/4に抑えられるという。デモには菱洋エレクトロが技術協力した。本誌が撮影。

同社がAIの推論実行時に演算器の使用率を計測したところ、GPUは20～30%、aiWareは約90%と大きな差が生じた（図5）。このことから、aiWareは「同じ処理速度なら、消費電力をGPUの1/3～1/4に低減できる」（Bialke氏）という。製造コストを決めるチップ面積もGPUに比べて1/3～1/4で済む。

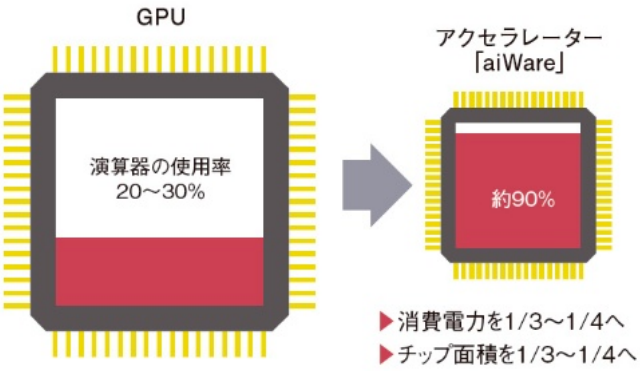

図5 演算器の使用率に差

AImotive社はさまざまな種類のニューラルネットで物体認識を行う際の演算器の使用率を計測した。その結果、GPUは20～30%だったのに対し、アクセラレーターは約90%と、GPUに比べて3～4倍高かったという。本誌が作成。

同社はaiWareの設計データをIP（Intellectual Property）として半導体ベンダーに提供する。2018年3月には中国半導体メーカーのVeriSilicon Holdings社と共同で2TMACS（Tera Multiply-Accumulate Operations per Second）の性能を持つテストチップを出荷する。2018年後半には別の半導体ベンダーと共同で20TMACSのテストチップを出荷する計画である。消費電力は1TMACS当たり1W以下を目指す。

数量によってはFPGAも選択肢

アクセラレーターは、SoCの中に組み込むと最も高い性能を実現できる。ただ、SoCは開発コストが高く、大量生産が見込める用途でしか使えない。AIを使う自動運転向けのSoCがどれほど大量に必要になるかは未知数である。数量が見込めない場合には、SoCではなくFPGA（Field Programmable Gate Array）が候補になる。FPGAは回路を自由に変えられるため、アクセラレーターとして使える。しかも標準品として大量生産されているため、少量の用途でも使いやすい。

日本アルテラ（インテル プログラマブル・ソリューションズ事業本部）はニューラルネットをFPGAで処理し、車両や歩行者を認識するデモを2017年11月に見せた（図6）。カメラで捉えたハイビジョン画質（720p）の映像に対し、30フレーム/秒の速度で4種類の物体を認識する。4種類の物体は、車両、歩行者、白線、他の車体が占有していない路面上の空間である。

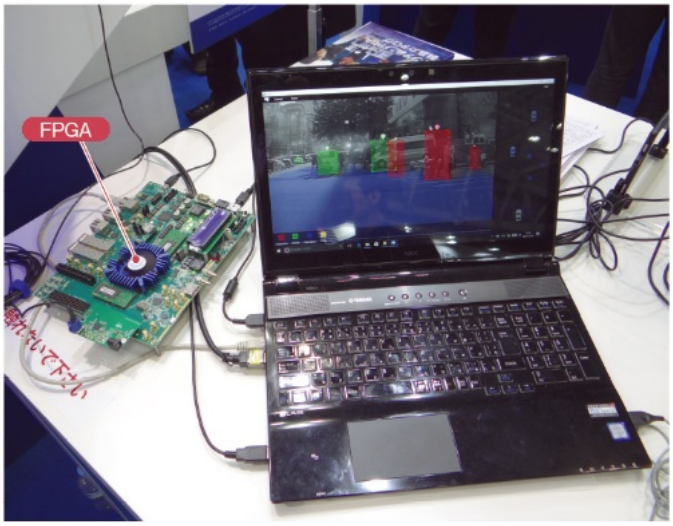

図6 FPGAで物体認識を低電力化

日本アルテラは車両や歩行者を低電力で認識するFPGAのデモを2017年11月の組み込みシステム技術展「Embedded Technology 2017」で見せた。本誌が撮影。

FPGAを使うことでニューラルネットの推論処理に必要な消費電力を8.5Wと「GPUに比べて半分以下にできた」（同社）という。FPGAはGPUに比べてメモリーアクセスの頻度を減らしやすく、メモリーの入出力回路で消費する電力を削減できる。FPGAには同社のミッドレンジ製品「Arria 10」を使い、ニューラルネットには中国Horizon Robotics社のアルゴリズムを使った。

速度5倍、消費電力1/10へ

FPGAと同等の機能を、より高速・低電力で実現する新型半導体の開発も始まった。太陽誘電と半導体設計会社のTRLが共同で開発している「MRLD（Memory-based Reconfigurable Logic Device）」は、FPGAに比べて動作速度を約5倍、論理回路の密度を約5倍にできるほか、同じ動作速度ならば消費電力を1/10以下に低減できる。

通常、論理回路はゲート（トランジスター）を組み合わせて、真理値（入力の「0」「1」に対する出力の「0」「1」の関係）を決めている。しかし、真理値をメモリーに書き込んで参照することでも同じような動作ができる。こうしたメモリーはルックアップテーブル（LUT）と呼ばれる。FPGAもMRLDも、LUTで論理機能を実現している点は変わらない。

ただ、FPGAはLUT同士を多数のスイッチを介して接続しているのに対し、MRLDはLUT同士を直接つないでいる点が異なる（図7）。MRLDは、FPGAのスイッチ部分で発生する信号遅延や電力消費をなくせる。また、スイッチ部分が専有するチップ面積も削減できる。「FPGAのスイッチ部分はチップ面積の約9割を占めるため、その効果は大きい」（TRL最高経営責任者の勝満徳氏）といえる。

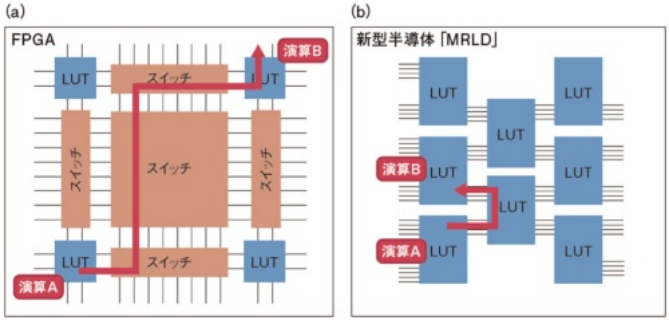

図7 スイッチを介さずに接続

（a）FPGAは、演算を担うLUTを多数のスイッチ（トランジスター）を介して接続するため、スイッチ部分で信号遅延や電力消費が発生しやすい。（b）これに対し、新型半導体の「MRLD」はLUTを直接つないだ構造であるため、高速、低消費電力化に向く。TRLの資料を基に本誌が作成。

MRLDは太陽誘電が特許を持ち、「他社のFPGA関連の特許には抵触しない」（同社営業本部新商品事業化企画担当部長の関口象一氏）という。チップの設計や販売はTRLが担当する。MRLDは従来のIC設計手法がほぼそのまま使えるほか、現在開発中の「機能マッピング」と呼ぶ新しい設計手法を使うと、「設計期間を従来の約1/10に短縮できる」（勝氏）。機能マッピングを使えば、「数百WのGPUと同等の機能を数Wで実現できる可能性もある」（同氏）という。

MRLDの回路構成はニューラルネットの構造と似ており、この特徴を生かしてAIに応用できる可能性がある。こうしたAIへの応用は大学と共同で研究中である。このほか、GPUで学習させたニューラルネットをMRLD上で動かす研究にも着手しており、「早ければ、3年以内にMRLDを使った自動車向けのAIを実用化したい」（同氏）という。

自動運転の「判断」を担う

自動運転では主に歩行者や車両の認識処理にAIを活用するが、認識した結果に基づいて判断を行う部分にもAIを利用する可能性が高い。デンソー子会社で半導体IPを開発するエヌエスアイテクスは、こうした判断の処理に適した新型プロセッサー「DFP（データフロープロセッサー）」を開発している。判断に関するアルゴリズムは研究途上にあり、どのような処理が求められるのか明らかになっていない。ただ、同社によると、画像処理のような画一的な処理ではなく、複雑な処理を短時間でこなす必要があるという。

例えば、車両の前方に二輪車が突然飛び出し、緊急回避が必要になったとする（図8）。歩行者がいない右側に進路を変えた瞬間、前方からレジ袋が飛んできたとしても、「レジ袋なら衝突してもリスクが低い」と判断する能力がプロセッサーに求められるという。一連の動作を、時間を細かく区切って見ていくと、「リスクの大きさや進路の判断は刻々と変わっていく」（エヌエスアイテクス開発部部長の伊藤雅之氏）。このような複雑な処理を短時間で行うためには、DFPのような新しいチップが必要とする。

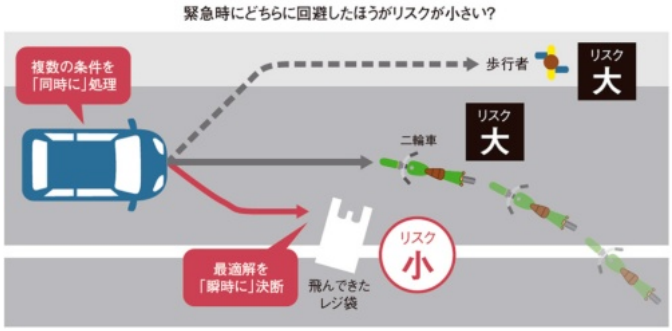

図8 複雑な判断を短時間で実行

レベル4以上の自動運転では、運転者に代わってシステムが瞬間的に複雑な判断を行う必要がある。例えば、前方から接近する二輪車を避ける際、経路の先にレジ袋が飛んできたとしても、「レジ袋なら衝突してもリスクが低い」と瞬間的に判断する能力が求められる。デンソーの資料を基に本誌が作成。

DFPは異なるデータ長のさまざまな演算に対応した演算器を多数備える。通常のプロセッサーは異なる種類の演算を順番通りに処理していくため、演算器の使用率が低くなる（図9）。そこでDFPでは、使用していない演算器に先々の処理を割り当て、演算器の使用率を高める。割り当てはDFP内部のハードウエアスケジューラーで行う。この技術は米ThinCI社が開発したもので、エヌエスアイテクスは同社と共同で車載半導体に仕上げる計画だ。

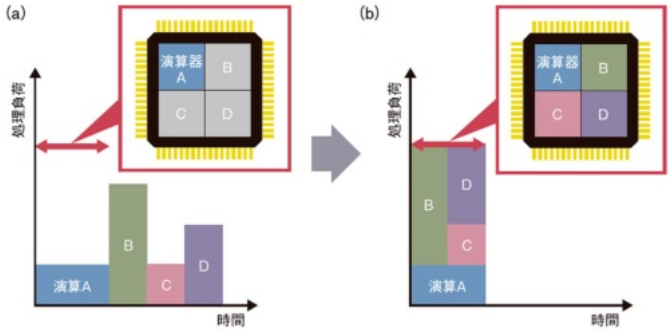

図9 演算器を効率良く使う

（a）通常のプロセッサーは、A～Dの演算を順番に処理するため、ある時間では演算器Aしか使っておらず、演算器B～Dは空いている。（b）DFPの場合、ハードウエアスケジューラーが演算の順番を適切に入れ替えて演算器A～Dを効率よく使えるようにする。デンソーの資料を基に本誌が作成。

1TOPS当たりの消費電力は約0.1WとGPUの1/10に低減することを目指す。1.6TOPSのDFPを最小構成とし、これを2個または4個並べて段階的に性能を増やせるようにする。2018年夏には最小構成の評価用テストチップを出荷する。現在、車載向けの半導体ベンダー数社と商談を進めており、2020年代前半に自動運転での実用化を目指す。

ニューラルネットの違いを吸収

AI市場に参入する半導体メーカーにとって追い風となる新規格もできつつある。米Khronos Groupはニューラルネットのファイル形式の違いを吸収する「NNEF（Neural Network Exchange Format）」と呼ぶ規格を策定中である（図10）。ニューラルネットにはさまざまなタイプがあり、その種類によって学習済みネットの出力ファイル形式が異なる。このため、半導体メーカーはこれらのファイル形式に対応した変換ツールを個別に用意する必要があった。

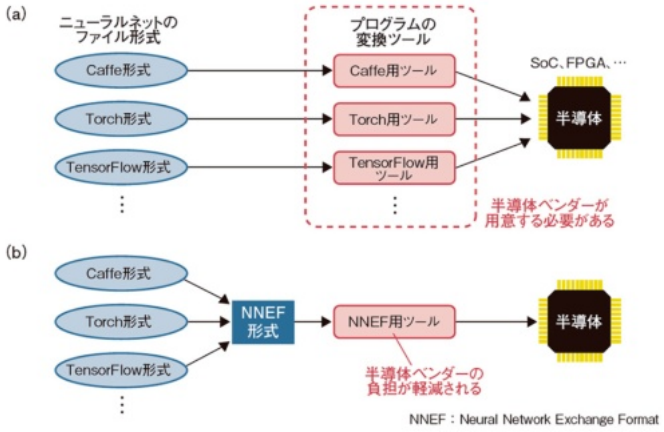

図10 半導体ベンダーの負担を軽減

（a）ニューラルネットのファイル形式はAIフレームワークの種類によって異なるため、これまでは半導体ベンダーが各AIフレームワークに対応した変換ツールを用意する必要があった。（b）ニューラルネットのファイル形式を統一規格の「NNEF形式」にできれば、変換ツールも1種類で済む。本誌が作成。

これに対し、Khronos Groupはさまざまなファイル形式のニューラルネットを統一フォーマットのNNEFに変換するツールを無償で提供する予定である。半導体メーカーはNNEF用の変換ツールさえ準備すれば、さまざまな種類のニューラルネットに対応できることになり、「AI半導体市場への参入障壁を下げられる」（NNEFの仕様策定に携わるAImotive社Lead AI Research EngineerのViktor Gyenes氏）。

NNEFの規格化を議論するワーキンググループには米Intel社をはじめ、多くの半導体メーカーが名を連ねている。2017年12月に暫定的な仕様である「NNEF 1.0」が公開される。その後、ワーキンググループ以外の業界の意見などを取り入れ、2018年中旬に正式版を出す。

NVIDIA社はさらに先を見すえる

このようにAI半導体の開発が活発化する中、先行するNVIDIA社はどう動くのか。「単純な価格競争には乗らないのではないか」との見方が多い。同社は高性能な製品をいち早く市場に提供し、先行者利益を上げる事業モデルを強みとする。同社は30TOPS、30WのDRIVE PX Xavierを提供するが、同等の性能でより消費電力が低く、安価な半導体が出てくるのは時間の問題だろう。その市場で価格競争が激化すれば、同社は次の市場に軸足を移すとみられる。

その兆候はすでに現れている。NVIDIA社はレベル5の自動運転に対応するため、性能を320TOPSとXavierに比べて1桁高めた車載AIコンピューター「DRIVE PX Pegasus」を2017年10月に発表した（図11）。同社が「ロボタクシー」と呼ぶ無人の自動運転車への利用を想定する。消費電力は500Wと大きいが、「320TOPSの性能が500Wで手に入るなら大歓迎」という顧客が多いという。レベル5の自動運転車は、配車などのサービスで収益を稼ぐ事業モデルのため、新サービスをいち早く立ち上げるために性能の高さを重視する顧客が多いようだ。NVIDIA社は次世代の市場を自ら創出することで他社との競争を優位に進めたい考えである。

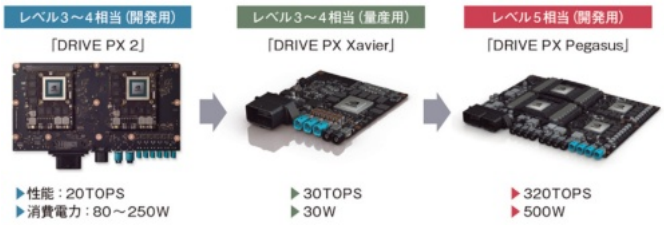

図11 NVIDIA社はレベル5に移行
レベル3～4の自動運転を見すえた半導体の開発競争が激しくなる中、NVIDIA社はレベル5に対応した「DRIVE PX Pegasus」を発表した。その性能は320TOPSと従来品に比べて1桁高い。NVIDIA社の資料を基に本誌が作成。

この記事のURL：
http://techon.nikkeibp.co.jp/atcl/mag/15/320404/120400065/