様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １５ 日現在

機関番号：１７４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０００８３

研究課題名（和文）深層学習向けニューラルネットワークチップの研究開発

研究課題名（英文）Neural network LSI for deep learning

研究代表者
尼崎　太樹（Amagasaki, Motoki）

熊本大学・大学院先端科学研究部（工）・准教授

研究者番号：５０４６７９７４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　本研究では，DNNの多種多様な用途・構造に対応するため再構成性を備えたAIアーキテクチャを研究開発し，低コスト，高ユーザビリティ，高速かつ超低消費電力なリコンフィギャラブルAIアクセラレータの回路構造を明らかにした．また，レイアウト設計レベルで性能評価を行った．AlexNetモデルを実装した際の動作周波数は350MHzであった．これらを加味した電力効率は883 [FPS/W]であった．本研究で開発されたアーキテクチャをベースに，MobileNetなどの新しいAIエッジコンピューティング向けのモデルへの対応を進める予定である．

研究成果の学術的意義や社会的意義
提案する深層学習チップの特徴は，NNの複雑さや規模に応じてHW構成を最適化できる点にある．特に，演算精度に着目してHW量を削減する回路最適化できる点が重要なポイントである．現在の深層学習は人工知能の先駆けに過ぎず，人間の知能に近づけるには膨大な計算量をいかに高速，低電力でできるかがカギとなる．一方，GPGPUや商用FPGAなどの汎用デバイスを用いたアプローチではこれらに対し限界が来るのは明らかである．提案する深層学習チップ開発を通して，IoTにおけるエッジサイドでの利用に対し、用途に合わせた最適化な形で処理を実行できる枠組みを示した．

研究成果の概要（英文）：In this research, in order to deal with a wide variety of applications and structures of DNNs, we have researched and developed an AI architecture with low cost, high usability, high speed, and ultra-low cost with reconfigurability. The circuit structure of a power-consuming reconfigurable AI accelerator is revealed. We also evaluated the performance at the layout design level using the standard cell library. Maximum operating frequency when implementing the pythorch-AlexNet model for CIFAR100 The frequency was 350 MHz. The processing power per second of the inference model is 100 [FPS], the power consumption is 0.11 [W], and estimated energy efficiency was 883 [FPS/W]. Based on the architecture developed in this research, we plan to work on models for edge computing.

研究分野：リコンフィギャラブルシステム

キーワード：AIチップ　ニューラルネットワーク

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人工知能分野で数十年来の歴史を持つ技術「ニューラルネットワーク（以下，NN）」が今，復権している．深い階層の NN を学習可能にする深層学習が登場したことで，画像処理や音声処理の分野で認識精度が飛躍的に向上している．深層学習では膨大な計算量を必要とするため，計算機によるサポート無しでは成り立たない．しかし，今後人間の大脳レベルの大規模 NN を想定した場合，計算機自体の消費電力が大きな障壁となり実現そのものが危ぶまれる．

NN は神経細胞に相当するニューロンとその接続のシナプス結合で構成される．深層学習で必要となる膨大な計算量に対し，これまではクラウドサーバや GPGPU を潤沢に投入する力技で速度向上を達成してきた．ところが，深層学習というパラダイムシフトが起きたにも関わらず，実装規模は従来の 10 倍しか向上していない．この理由は消費電力にある．GPGPU では速度向上の対価として大量の消費電力を消費している．今後，ニューロン数が約 106 倍となる人間の大脳相当規模の NN を想定した場合，これまで同様 GPGPU を投入できる可能性は限りなく低い．

一方で GPGPU と競合する手段として FPGA(Field Programmable Gate Array)がある．FPGA は速度性能では GPGPU に劣るが，ワットあたりの処理性能（電力効率）は GPGPU よりも高い．ただし，大半の研究では学習で得られたパラメータ(ニューロン数やシナプス結合数）を保ちつつ，演算ビットの精度を削ることで単一チップ内に収めるよう設計されている．このため，パラメータに依存した回路設計が行われており汎用的でない． また，GPGPU と同様に今後 NN の規模が増加した場合の設計方法が確立されていない．このほか GPGPU や FPGA のアプローチとは別に，大規模 NN の実装を目的とした脳型コンピュータの研究が進められているが実現には至っていない．

２．研究の目的

GPGPU など高い処理能力や大規模メモリを前提とした従来の計算処理体系とは異なり，NN の特徴に特化した専用チップを開発し，高性能・低消費電力な深層学習向けアクセラレーション手法を実現することを目的とする．研究期間内に以下の 3 項目を明らかにする

・柔軟性と拡張性を備えた NN 専用アーキテクチャ
・データ駆動型に対応した低消費電力回路方式
・大規模 NN に対応した LSI 設計方式の決定

第 1 に演算精度（演算ビット数，シナプス結合数など）の最適化を行う．その上で，NN に応じて最適な性能・機能・規模を選択できる柔軟性と拡張性を備えたアーキテクチャを開発する．第 2 に負荷に応じて動作する回路量を変えることで低消費電力化を目指す．必要のないセルはパワーゲーティング等を駆使して電力を抑え，必要最低限の回路のみ動作させる．第 3 にデバイスレベルのアプローチで電力効率の最適化を図る．最終的に従来手法（GPGPU，FPGA）との比較を行い，提案手法の優位性を示す．

３．研究の方法

本研究では，GPGPU と匹敵する回路性能をもち，FPGA より電力効率が高い NN 専用チップの実現を目的とする．期間は 3 年とし，以下の 3 項目に関し研究を行う．①ニューラルネットワーク専用アーキテクチャの研究，②イベント駆動型低消費電力方式の研究，③3 次元積層方式を用いた LSI 化の研究である．初年度は①にて NN に最適化された低精度演算セル方式を検討し，②では負荷に応じて演算量を切り替えるイベント駆動型の低消費電力方式の検討を行う．①と②にて回路レベル最適化を行い基本アーキテクチャの仕様を決定する．これに基づき③では 3 次元積層方式を用いたデバイスレベル最適化のアプローチをとる．レイアウト設計までを行った上で，GPGPU と FPGA との比較を行い優位性を示す．

４．研究成果

４．１　FPGA アクセラレータ

４．１．１　2 のべき乗演算を用いた FPGA アクセラレータ

CNN の組込み機器への実装を考えた場合，低消費電力かつ高速な処理が可能である FPGA（Field Programmable Gate Array）は有望な選択肢となる．しかしながら，FPGA に CNN を実装する際は，内部で膨大な回数実行される積和演算回路の構成および重みを読み込む際のメモリアクセスについて工夫する必要がある．そこで本研究では，CNN の重みを 2 のべき乗に近似する手法を提案する．これにより，積和演算回路における乗算はシフト演算に置き換え可能となる．また，重みを近似する際は CNN に再学習を施すことで認識率の低下を抑制し，近似後は閾値以下の重みをプルーニングすることで重みの表現に必要なビット幅が期待できる．

重みの 2 のべき乗近似による認識率の低下を抑制するため，CNN に対して再学習を適用する．なお，再学習はソフトウェア上で実行し，浮動小数点形式のデータを用いて演算を行う．まずはじめに，CNN が持つオリジナルの重みを 2 のべき乗化させたうえで順伝播処理を実行する．次に誤差逆伝播処理を実行し，正解データとの誤差が小さくなる方向へ重みを更新する．このとき，

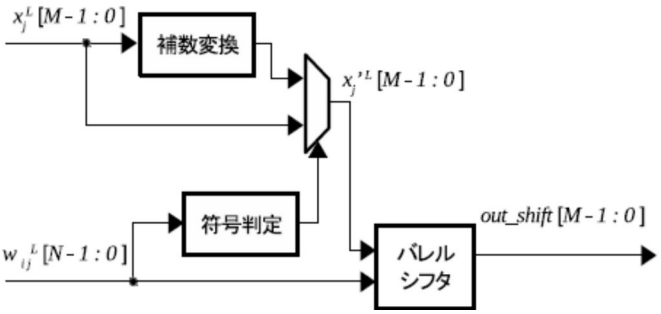

図 1　乗算演算回路

更新後の重みは 2 のべき乗に限定されない値をとることになる．更新後の重みを再び 2 のべき乗化したのち，学習が収束するまで上記の処理を繰り返す．

図 1 に，シフト演算を用いた乗算回路の構成を示す．入力 xjL は M ビット固定幅の信号であり，入力 wLij は N ビット幅の信号であるとする．なお，CNN では固定小数点による計算は演算精度への影響が小さいため，これらの数値は固定小数点で表現する．また，重みをメモリに保存する際は実際の値ではなく，シフト幅ごとに割り当てられたラベルを格納する．このとき，重みの最下位ビットは 1 ビットシフトを実行するかどうかを判断するためのセレクト信号として用い，その次のビットはさらに 2 ビットシフトを実行するかどうかを判断するためのセレクト信号として用いる．これは，バレルシフタ回路におけるシフト幅の表現方法に即したものとなっている．また，重みの最上位ビットは重みの符号を表現するためのビットとして割り当てる．さらに，ラベル化の際は重みの値が 0 である場合を表現するため，ラベルを 1 つ余分に用意する．

本評価では，ベースラインとして BVLC（Berkeley Visionand Learning Center）によって公開されている AlexNet の学習済みモデルを利用し，シミュレーションはディープラーニング設計フレームワークの Chainer によって実行する．Chainer では，CNN モデルの重みは数値計算ライブラリである numpy の ndarray 型多次元配列として表現される．また，配列の要素値は単精度浮動小数点の形式で保存されている．CNN の重みを 2 のべき乗に近似する際，C 言語で作成した Python ライブラリを用いて，重みが格納されているメモリの値をビット単位で直接変更する．単精度浮動小数点の指数部および仮数部をそれぞれ操作することで，オリジナルの値との差が最も小さくなるような 2 のべき乗値に重みが変換される．FPGA に実装する回路については C++によって記述し，XILINX VivadoHLS 2016.04 を用いて高位合成を行う．また，ビットストリームの生成には XILINX Vivado 2016.04 を使用する．論理合成の際のターゲットデバイスは，Virtex7 XC7VX485T が搭載された VC707 を設定する．また，入力画像は符号部 1 ビット，整数部 8 ビット，小数部 7 ビットの固定小数点形式のデータとして扱う．

表 1　重みの 2 のべき乗近似による認識率の変化

2 のべき乗近似の適用範囲	認識率（%）	
	top-1	top-5
ベースライン	57.00	80.14
畳込み層 1 から畳込み層 5（再学習なし）	51.22	75.25
全結合層 6 から全結合層 8（再学習なし）	56.68	79.93
全体（再学習なし）	50.76	75.07
全体（再学習あり）	56.72	79.82

AlexNet の重みを 2 のべき乗に近似した際の認識率の変化について表 1 に示す．なお，再学習を行う際は損失関数として Softmax Cross Entropy を用い，最適化関数として MomentumSGD を用いた．，再学習を適用することによって，AlexNet 全体の重みを 2 のべき乗した際の認識率の低下を約 0.3%に抑えることができた．また．再学習時の損失の変化をみても，重みを 2 のべき乗に近似した場合でも学習が正常に進められることがわかった．

４．１．２　高速光通信を用いた CNN 分割実装

続いて，CNN をマルチ FPGA システムに実装することで低消費電力かつスケーラブルなシステムを実現した．マルチ FPGA システムでは FPGA 間の通信がボトルネックとなるが高速シリアル光通信を用いることで解決を図った．　本研究ではエミュレーション周波数の低下，I/O ピンの不足，複数の FPGA 間の配線が困難である点について，高速シリアル光通信を使用することで解決を図る．光通信は Virtex-5 までの電気通信に比べて 10Gbps 以上の高速な通信や長距離の通信が可能である．提案手法にといて，まず畳み込み層や隠れ層の重みを各 FPGA に分割して保存する．畳み込み層の出力をそれぞれお互いの FPGA に高速シリアル光通信を用いて送信する．そして，隠れ層で同じ FPGA から送られる値とほかの FPGA から送られてきた値と重みとの乗算を行う．そして各ニューロンで和を計算して出力層に送っていく．その際に再びその値を各 FPGA に送信する．この分割方法のメリットは非常にシンプルである点である．そのため複雑な送信を行うことがなく，複数の FPGA に容易に分割して実装が期待できる．

本研究ではエミュレーション周波数の低下，I/O ピンの不足，複数の FPGA 間の配線が困難である点について，高速シリアル光通信を使用することで解決を図る．光通信は Virtex-5 までの電気通信に比べて 10Gbps 以上の高速な通信や長距離の通信が可能である．ここでは，畳み込み層や隠れ層の重みを各 FPGA に分割して保存する．畳み込み層の出力をそれぞれお互いの FPGA に高速シリアル光通信を用いて送信する．そして，隠れ層で同じ FPGA から送られる値とほかの FPGA から送られてきた値と重みとの乗算を行う．そして各ニューロンで和を計算して出力層に

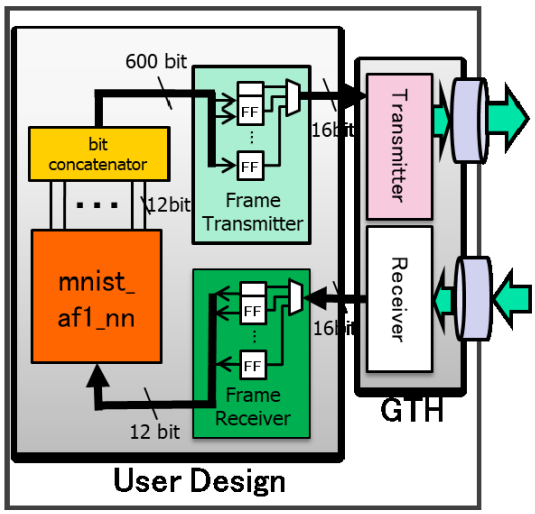

図２　高速シリアル通信を用いたシステム

送っていく．その際に再びその値を各 FPGA に送信する．この分割方法のメリットは非常にシンプルである点である．そのため複雑な送信を行うことがなく，複数の FPGA に容易に分割して実装することができる．図 2 に提案システムの概要を示す．この図では隠れ層の IP をシステムに実装している．mnist_af1_nn は生成した隠れ層の IP で ap_ctrl_hs で制御するが，この図では省略している．この IP の入力はプーリング層からの送られてくる 12bit の 1 本の信号である．出力も 12bit ではあるが層のニューロン数で並列化を行っているため信号は 50 本となる．その出力を bitconcatenetor で 600bit の 1 本の信号とし，Frame Transmitter に送る．Frame Transmitter でトランシーバへトランシーバのポート幅である 16bit ずつ送信する．そして光ケーブルを通してほかの FPGA へ値を送信する．本実験では Xilinx 社の KintexUltraScale KCU105 評価ボードを用いる．この KCU105 には Xilinx 社製 FPGA である Kintex XCKU040-2FFVA1156E が搭載されている．I/O 端子数は 520 である．また，このボードには高速シリアルトランシーバである GTH トランシーバが 20 個搭載されている．開発ツールは Vivado Design Suite 2016.2 を用いた．

これを元に，分割前の従来のシステムと 2 分割を行った提案システムのリソース使用率を評価した．全体的に分割前と分割後ではリソース使用率はおよそ半分ほどに減少している．最も BRAM と DSP を使用している隠れ層でも BRAM は約 0.6 倍，DSP は 0.5 倍になっており，大規模回路を実装する場合でも回路を分割することで実装可能となることがわかる．出力層の BRAM の使用率が 0%となっているのは FF や LUT で保存可能なほど重みの数が少ないためであると考えられる．続いて分割前と分割後の各層の IP と転送時間のレイテンシについて述べる．今回，同軸ケーブル上の遅延は考慮せずに転送のレイテンシを算出している．高速シリアル光通信を用いてほかの FPGA に値を送るのは畳み込み層と隠れ層である．分割前と分割後の IP のレイテンシを比べると，畳み込み層で約 0.7 倍，全結合層で約 0.8 倍となっており，分割後の方がレイテンシが少なくなっている．これは重みの数が半分に減ったことで 1 つの FPGA における計算量が削減されたためと考えられる．

４．２　重み転送専用ラインをもつ専用 LSI

4.1 で得られた知見を基に，専用 LSI を開発した．本チップは入力用，重み用，コントローラ用の 3 種類の SRAM を持つ．SRAM のサイズは各 512kB，全体で約 1.5MB である．中心には MAC 演算器と配線部からなる PB(Processing Block)が正方形になるように並べられている．PB の数は 64 列×64 行で合計 4,096 個存在する．PB はそれぞれ上下左右の PB と接続されており，この接続を再構成することでリコンフィギャラビリティを確保している．また，コントローラはマイクロコードで制御されており，層の処理に合わせて制御信号を生成することが可能である．

畳込みニューラルネットワーク（以下 CNN）における畳込み層での処理は，物体認識においてその特徴抽出を行う際に重要な役割を果たし，代表的な CNN である AlexNet では全 8 層のうち 6 層を畳込み層が締めている．実際に CNN の計算量に着目すると，畳み込みフィルタ演算部がそのほとんどを占めており，これまでは計算アルゴリズムの改善に重点が置かれてきた．しかし，CNN モデルの複雑化に伴い，フィルタ演算に必要なデータへのメモリアクセス回数増加が速度性能低下の大きな要因となっている．

畳込み層のフィルタ演算は画像処理と同じように積和（以下 MAC）演算回路を用いて計算される．注目すべきは，畳込み層では出力チャネル毎に畳み込みフィルタを備えるが，この畳込みフィルタは入力チャネル全てに同じものが適用される点である．これは，アクティベーションデータを用いて各特徴マップの出力を求める際に再利用できることを意味する．

本研究では，畳込みフィルタ係数とアクティベーションデータがメモリに格納されたシステムにおいて，畳込み層の特徴を利用してメモリアクセスを大幅に削減できる方式，およびその回路構成について提案を行った．演算処理を効率的に行うために，メモリからロードされたアクティベーションデータに対し，演算部でそれらのデータを使いまわすことでモリアクセスを大幅に削減する．ここでは積和演算回路（PE）と近傍接続を要する配線部を 1 つの処理ブロック（PB）とし，それを 2 次元格子状に並べた PB アレイで畳込み部の計算を並列に行う．アクティベーションと畳み込みフィルタのデータはメモリに格納され，これらはコントローラにより制御される．ここでは当該メモリを SRAM としており，SRAM の容量を超えるデータは外部の DRAM とやりとりすることを想定しているが，実現方法はこの限りではない．メモリと PB アレイ間の接続は専用配線をもち，これらのデータは PB の配線部を通して積和演算回路へと入力さ

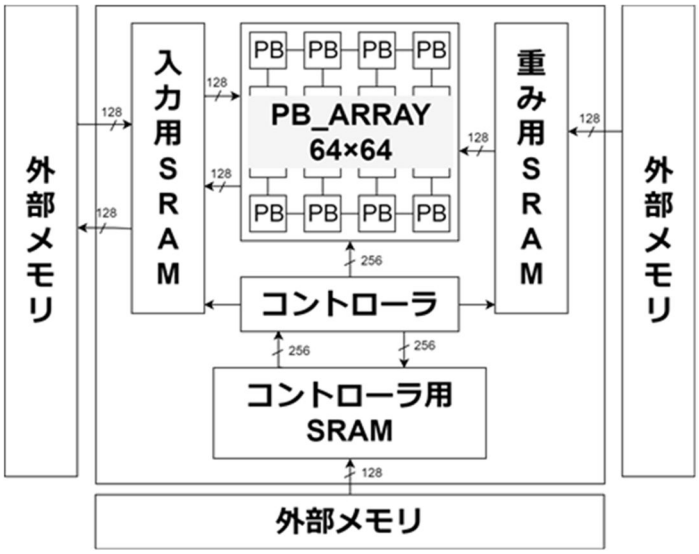

図３　ニューラルネットワーク LSI

れる．また，積和演算回路の計算結果をメモリに格納する際も配線部を通して行われる．加えて，メモリへのアクセス回数を削減するため，アクティベーションデータを移動させる機構を備える．DNN モデルの実装については，入力チャネルの行レベルのデータを一斉にメモリから PB アレイへロードし，並列に演算を行う．出力時も出力チャネルにおいて行単位で PB アレイの計算結果をメモリへ書き戻す．

図 1 に力チャネル＝64，PB アレイ数 64×64 の畳み込み演算の例を示す．最初に PB アレイの行には同入力チャネル内のアクティベーションデータが格納されている．重みの専用ラインを通してメモリから 64 個分の重み情報がロードされ積和演算を行う．データ再利用のため，行方向にアクティベーションデータのシフトを行う．これを入力特徴マップの行数分繰り返し行う．この後，重み関数が移動する際は左シフトして同じ演算を繰り返す．

表２　処理時間

層名	処理時間 [ms]	
	オリジナル版	50 % プルーニング版
第 1 層	0.02	0.02
第 2 層	0.67	0.67
第 3 層	0.48	0.48
第 4 層	0.65	0.65
第 5 層	0.43	0.43
第 6 層	9.89	4.99
第 7 層	9.89	4.99
第 8 層	2.54	1.32
合計	24.58	13.56

Pytorch 版 AlexNet を処理した場合の処理時間を表 2 に表す．また，オリジナル版と 50 % プルーニング版の認識精度を表 3 に表す．表 2 より，処理時間の大半を全結合層である第 6，7，8 層が占めていることが分かる．これは，畳み込み層は重みの転送は 1 サイクルで終わるため，演算時間が処理時間の大半を占める．対して，全結合層は重みの転送に複数サイクルかかるため，転送時間が処理時間の多くを占めるからである．しかし，畳み込み層の処理が高速に行えているためオリジナル版でも 24.58 ms と 30 FPS を達成している．また，50 % プルーニング版では，全結合層の処理時間がほぼ半減しており，プルーニングによって処理時間を短縮できている．そのため，処理時間は 60 FPS を達成している．今回は再学習を行わずにプルーニングのみを行ったため，プルーニングを 50 % に抑えた．再学習を行う場合はより多くの重みをプルーニングでるため，更なる処理時間短縮を見込める．よって，再学習も併用した場合，120 FPS 以上の処理速度も達成可能だと考えられる．

表３　認識精度

モデル	認識精度 [%]
ベースライン (32 bit float)	63.43
オリジナル版 (8 bit int)	62.03
50 % プルーニング版 (8 bit int)	61.57

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 千竈純太郎・中原康宏・尼崎太樹・久我守弘・飯田全広・末吉敏則
２．発表標題 高速シリアル光通信を用いたCNN分割実装の検討
３．学会等名 電子情報通信学会
４．発表年 2018年

１．発表者名 Theingi Myint（Kumamoto）・Qian Zhao（Kyutech）・Motoki Amagasaki・Masahiro Iida・Toshinori Sueyoshi
２．発表標題 Resources Utilization of Fine-grained Overlay Architecture
３．学会等名 電子情報通信学会
４．発表年 2018年

１．発表者名 中原康宏・千竈純太郎・尼崎太樹・飯田全広・久我守弘・末吉敏則
２．発表標題 CNN accerarator using power-of-two weight and pruning
３．学会等名 電気・情報関係学会九州支部連合大会
４．発表年 2018年

１．発表者名 千竈純太郎・中原康宏・尼崎太樹・飯田全広・久我守弘・末吉敏則
２．発表標題 CNN implementation using High Speed Optical Serial Links
３．学会等名 電気・情報関係学会九州支部連合大会
４．発表年 2018年

１．発表者名
宇都宮誉博，尼崎太樹，飯田全広，久我守弘，末吉敏則

２．発表標題
重みの2のべき乗近似を用いたCNNのFPGA実装に関する一検討

３．学会等名
電子情報通信学会

４．発表年
2017年

１．発表者名
宇都宮誉博，尼崎太樹，飯田全広，久我守弘，末吉敏則

２．発表標題
2のべき乗近似とプルーニングを用いたCNN向けFPGAアクセラレータ

３．学会等名
電子情報通信学会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Arch研究ホームページ
www.arch.cs.kumamoto-u.ac.jp
熊本大学コンピュータアーキテクチャ研究室
www.arch.cs.kumamoto-u.ac.jp

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考