卒　業　論　文

題　　目

マルチFPGAボード上でのGoogLeNetの実装に関する研究

年　　度

平成29年度

所　　属

慶應義塾大学

理工学部　情報工学科

指導教員

天野 英晴 教授

氏　　名

61401007

飯塚　健介

卒 業 論 文 要 旨

学 科	情報工学	学 籍 番 号	61401007	フリガナ 氏 名	イイヅカ ケンスケ 飯塚 健介

(論 文 題 名)

マルチ FPGA ボード上での GoogLeNet の実装に関する研究

(内 容 の 要 旨)

昨今，人工知能が最新技術のトレンドとして，様々なメディアに取り上げられ様々な製品に組み込まれている．特に画像認識や音声処理，物体検出などの分野で大きな貢献を果たしているニューラルネットワークは一躍注目されていて，研究開発が盛んに行われている．ニューラルネットワークの一種である畳み込み演算を主な計算とする畳み込みニューラルネットワーク (CNN) は精度向上のために計算量が増加する傾向にある．CNN の認識精度向上のためには高速化，電力性能向上が求められている．しかし，汎用プロセッサではその要求を満たすことができないので，各半導体メーカや研究機関は専用のアクセラレータの開発に取り組んでいる．日本でも国立研究開発法人新エネルギー・産業技術開発機構 (NEDO) は「省電力 AI エンジンと異種エンジン統合クラウドによる人工知能プラットフォーム」と銘打ったプロジェクトで複数の FPGA，GPU，メモリなどの異種ノードを多数接続した大規模計算基盤 Flow-in-Clowd(FiC) を開発している．複数の FPGA は高機能スイッチノードとして多数の高速リンクが接続され，FiC の高速通信のスイッチングの役割を担う．このマルチ FPGA システムは主演算を行う GPU ノード更に AI エンジンとしての役割も担う．本研究ではマルチ FPGA システムに 2014 年の ILSVRC で最高精度をマークした CNN モデルの 1 つである GoogLeNet を実装し, 性能で CPU の○○倍，GPU の○○倍を達成し電力効率で CPU の○○倍，GPU の○○倍を達成した．

（内容の要旨は約25行程度で記入のこと）

目 次

図 目 次

表 目 次

第1章

序論

1.1 本研究の背景

人工知能と称される機械学習をベースにした技術は爆発的な普及を見せていて日夜，メディアで取り上げられるだけでなく，自動運転やスマートスピーカー，スマートフォン向けアプリケーションなど様々なシステムに取り込まれている．しかし，人工知能のさらなる普及にはその計算基盤が必要である．その中でも特に画像認識や物体検出などの分野で活躍する畳み込みニューラルネットワーク (CNN) はその計算の特性から汎用 CPU では効率よく演算処理ができない．インテルや NVDIA など大手半導体メーカーを始めとして Google や Micrsoft なども人工知能向け専用アクセラレータの開発に心血を注いでいる．各社，研究機関は GPU，ASIC，FPGA など様々なデバイス，手法で高速化を図る．その中でも FPGA はその電力効率のよさ，開発周期の短さ，再構成可能であることから注目され研究がなされている．日本でも国立研究開発法人新エネルギー・産業技術開発機構 (NEDO) は「省電力 AI エンジンと異種エンジン統合クラウドによる人工知能プラットフォーム」と銘打ったプロジェクトで複数の FPGA，GPU，メモリなどの異種ノードを多数接続した大規模人工知能計算基盤 Flow-in-Clowd(FiC) を開発している．この FiC はデータセンターなどに導入されるクラウドシステムである. 主演算装置となる複数の GPU を複数の FPGA のスイッチノードに接続し，高速通信を行う．高機能スイッチノードととなるマルチ FPGA は多数の高速リンクが接続され，FiC の高速通信のスイッチングの役割を担う．さらにこのマルチ FPGA システムはスイッチノードという役割に加え，AI エンジンとしての役割も担う．そこで本研究ではマルチ FPGA システムの試作ボードである FiC-SW1 を複数枚用いて，CNN のモデルである GoogLeNet を実装し，評価を取った．

1.2 研究目的

本研究の目的はマルチ FPGA システム上に GoogLeNet を実装し，既存研究や汎用 CPU，GPU に対して性能向上を目指すことである

1.3 本論文の構成

??章では実装対象である GoogLeNet と畳込みニューラルネットワークの概要を説明す

る．**??**章では本研究で用いるマルチ FPGA システムとそのプロジェクトの概要を紹介する．**??**章では本研究に関連する先行研究について説明する．**??**章では GoogLeNet の並列化手法について説明する．**??**章では**??**での並列化を考慮した実装方法について説明する．**??**章では本研究の評価を行う．**??**では本論文の結論を述べる．

第2章

GoogLeNet

本研究でマルチ FPGA 上に実装するアプリケーションである GoogLeNet について説明する．GoogleNet は畳込みニューラルネットワーク (CNN: Convolutional Neural Network) のモデルの 1 つである．まず CNN の基本的な演算について説明する

2.1 Convolutional Neural Network

2012 年の ILSVRC() で登場した AlexNet により CNN は特に画像認識や物体検出の分野で優れた識別精度をマークしたことから世界で注目されるようになった,

2.1.1 Neural Network

ニューラルネットワークは動物の神経ニューロンが接続され，神経物質が伝搬されるように演算モジュールを層として複数，結合していく．演算ネットワークでは神経物質の代わりに例えば画像における画素値のような入力ベクトルの演算結果を伝搬していく．ニューラルネットワーク及び機械学習では２つの演算フェーズ，学習と推論がある．学習では教師データ (識別された画像) をもとに各演算層のパラメータを決定する，推論では学習で得たパラメータを用いて，入力値 (未識別の画像) から演算 (入力画像の識別) を行う．本研究では推論アルゴリズムにフォーカスしたアクセラレータを実装するので推論演算で用いられるアルゴリズムについて説明する

2.1.2 Convolution

CNN はその名前にも含まれているように式で表される畳み込み演算と呼ばれる行列の積和演算がその主なアルゴリズムである．CNN では各層の演算結果を次の層の入力値として受け渡す．これを特徴マップと呼ぶ．特徴マップは畳み込み層や全結合層で重みフィルタと呼ばれる学習結果から得られる行列と積和演算が行われる式で畳み込み演算，マックスプーリング，全結合の演算を示す

$$output(x,y)^i = input(x,y)^i / (k + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a^j_{x,y})^2)^\beta \tag{2.1}$$

$$output(x,y)^i = max(input(i,j)|x-k <= i <= x+k, y-k <= j <= y+k) \tag{2.2}$$

$$output(i) = bias(i) + \sum_{j=0}^{N} weight(i,j) * input(j) \tag{2.3}$$

これらの演算を各層で処理し伝搬していき出力されるという構造をとっている．CNN では多くのネットワークモデルでこの畳込み層が主である．

2.2 GoogLeNet

GoogLeNet は 2014 年の ILSVRC で最高精度をマークした CNN のモデルの一つである．構成要素は上述の畳み込み演算，プーリング処理などが主であるという点においては AlexNet などと違いはないが，それぞれの層の接続の仕方，層の深さが異なる，AlexNet, GoogLeNet それぞれについて図にその全体像を示す

両者を比較すると，GoogLeNet のほうが，層がより深くなっていること，さらに横に広がっていることがわかる. GoogLeNet は層を深くする代わりにそのフィルタサイズを小さくすることで，計算量，メモリアクセスを減らすように設計されている．○○の研究によると一つの大きなフィルタによる畳込みよりも複数の小さなフィルタによる畳込みのほうがより高い精度を出すことができるとされている．さらに GoogLeNet では横に広がった複数の層を Inception 層と名付け，これを複数層重ねる設計をしている．

2.3 Inception

Inception 層は次の図で示される (これは図の一部を拡大したものと同じである)

これを見ると横に広がった層の中で 1 ＊ 1conv と記されている層がある. この層ではサイズ１のフィルターを用いて，畳込み演算を行っている. この層は，次元削減を行っている．これは入力チャネル (入力行列の深さ) に対して少ない層のフィルタを畳み込み演算することでその深さを削減する，これによって計算量が減るだけでなく，精度向上が実現できる．Inception 層の出力の手前の DepthConcat 層は横に広がった層のそれぞれの出力結果を図○のように深さ方向に結合することで一つの行列として出力値にまとめる処理を行っている．この特徴的な Inception 層を積層していくことで，GoogLeNet は構成される．表 **??**に GoogLeNet での各層で必要なパラメータ数をまとめる．

この表を見ると AlexNet に見られるような全結合層がないことがわかる．GoogLeNet では高い認識精度を保ったまま，計算量の多い全結合層を削除することで全体の計算量を減らしている．

表 2.1: GoogLeNet における各層の構成

type	patch size/ stride	output size	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64							2.7K	34M
max pool	3×3/2	56×56×64								
convolution	3×3/1	56×56×192		64	192				112K	360M
max pool	3×3/2	28×28×192								
inception (3a)		28×28×256	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480								
inception (4a)		14×14×512	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832								
inception (5a)		7×7×832	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024								
dropout (40%)		1×1×1024								
linear		1×1×1000							1000K	1M
softmax		1×1×1000								

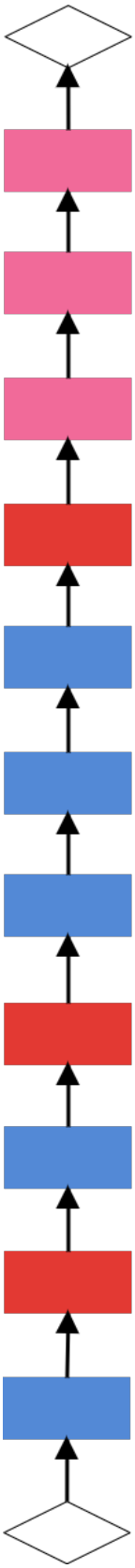

図 2.1: AlexNet のアーキテクチャ

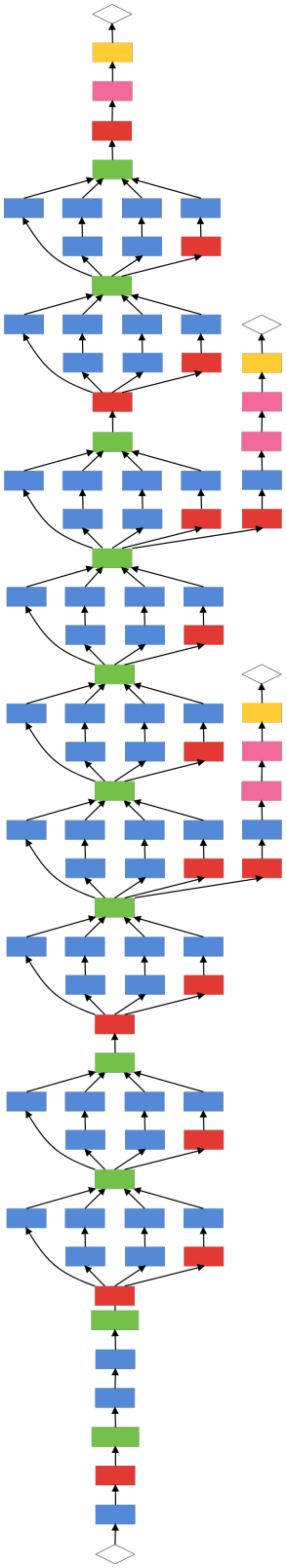

図 2.2: GoogLeNet のアーキテクチャ

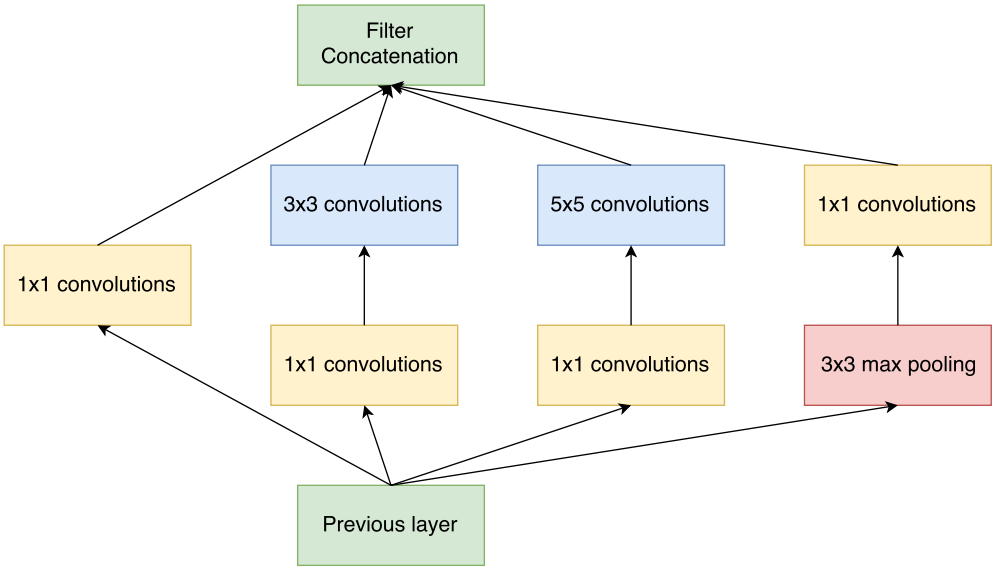

図 2.3: Inception 層

第3章

FiCSW

本章では，FiC システムとマルチ FPGA システムである FiCSW の概要を説明する．

3.1 FiC の概要

FiC は NEDO のプロジェクトとして以下のようなシステムの構成を目指している１，様々な種類の処理装置を適材適所で組み合わせて運用する「異種エンジンシステム」とすることで，汎用プロセッサを遥かに超える演算能力を有する２，光通信技術の導入により，高速通信が可能なインターコネクトを実現することでホストプロセッサを介することなくエンジン同士を接続し通信のオーバーヘッドの小さなシステムを構築すること

人工知能計算基盤としてのクラウドシステムなので，大量データによる学習フェーズなど即時性を必要としない演算処理を担うことが想定される，それに加え，様々なアプリケーションで柔軟にエンジンを効率よく割り当てるためにエンジン間を動的に変更することが可能なネットワークを構成する必要がある．

3.1.1 FiC のアーキテクチャ

FiC は３つの基板 (ノード) とそれらをつなぐネットワークから構成される．ノードには本研究で用いる FPGA ノードだけでなく GPU ノード，メモリノードがある．これらは図**??**に示すような接続により計算クラスタとして各種アプリケーションに用いられる

FPGA はシステムにおいて高機能スイッチノードとしての役割が期待され。他の FPGA や GPU，メモリノードと接続されることで GPU 間の高速通信を実現するとともに FPGA にプログラムされたデータ処理機構などを組み込むことでホストプロセッサの介在を不要とする．メモリノードは FPGA ノードに接続されることで，個々の GPU に大容量 DRAM をもたせる必要がなくなり，GPU デバイスへのメモリコピーをなくすことができ，メモリの利用効率を向上させることができる．

3.2 FiC-SW の概要

FiC において高機能スイッチとなる FPGA ボードは FiCSW と名付けられ，その試作ボードは図**??**である．

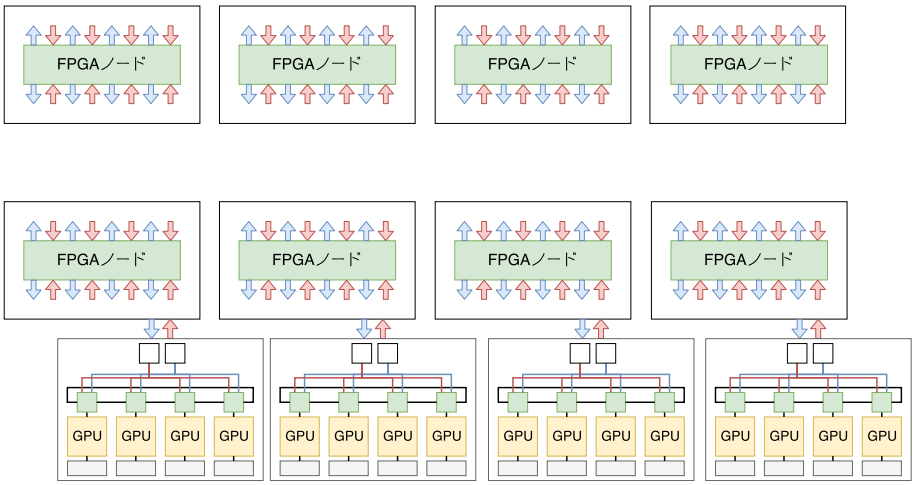

図 3.1: FiC のアーキテクチャ

図**??**中に赤枠で囲まれた 4 つの要素で構成されている。図の注釈の順に 1，Xilinx 社製の UltraScale XCKU095-FFVB2104 2, 高速シリアルリンク 8 チャネル，32 リンク 3,16GB の DRAM2 バンク 4,Raspberry Pie3 を搭載している。Raspberry Pie3 を用いることで，FPGA を遠隔操作して動的に再構成することができる．

3.3 FiC-SW での通信様式

FiC では将来的に広帯域光通信技術をシステムの相互接続ネットワークに用いることが想定されている．光信号のままスイッチすることで数十 Tbps の性能をもつインターコネクトをコストを抑えながらクラウド内で利用することができる．光信号はサーキットスイッチを用いることが現実的である．今回のボードは従来の電気信号による通信を行うが来たる光通信の将来を想定して，サーキットスイッチによるネットワークを構成する．サーキット数を増やすために時分割多重 (TDM: Time Division Mutipling) による通信様式をとる．

図**??**は 4 リンクからなるシリアル入力から同じく 4 リンクからなるシリアル出力へのスイッチの模式図である。各リンクが 4 スロットを有していると仮定している。図中に示される入力スロットと出力スロットの接続が回線を確立する。図の破線矢印で示されるように 1 スロットから複数スロットへの出力を設定することも可能なので、容易にブロードキャストを行える。

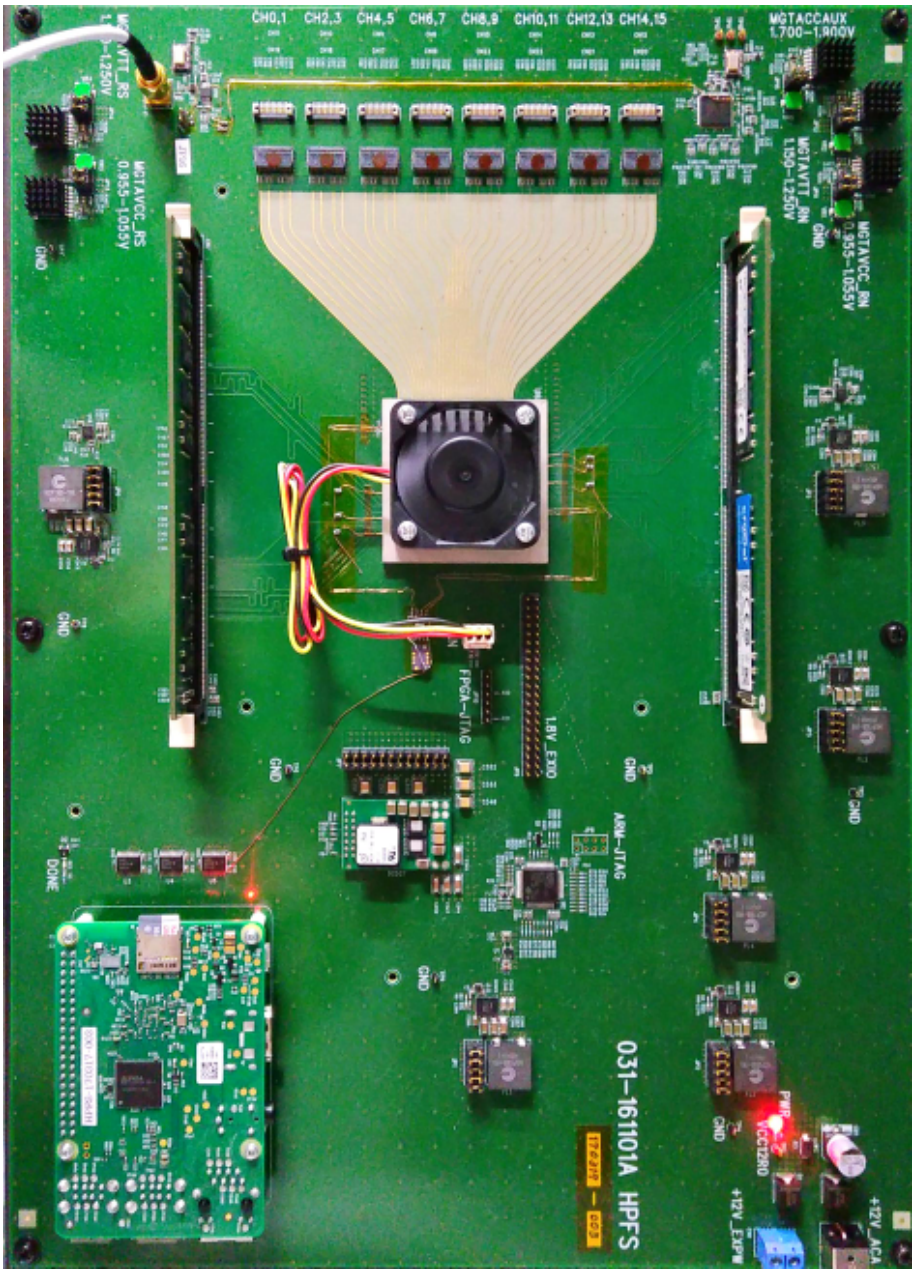

図 3.2: FiC-SW の試作ボード

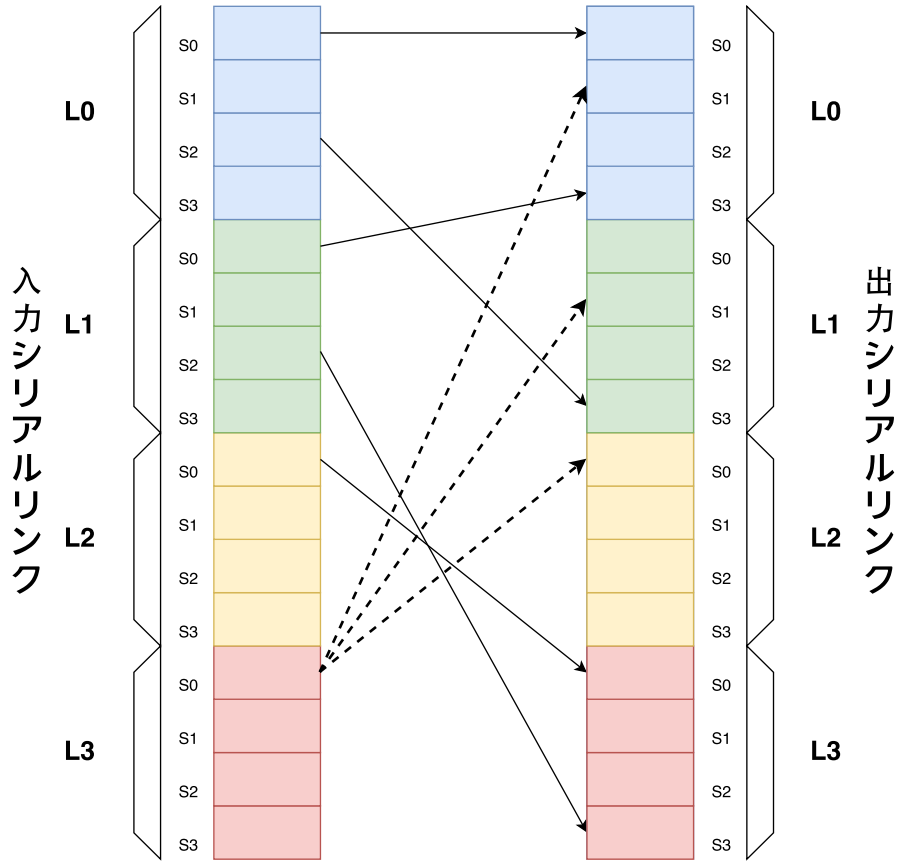

図 3.3: FiC のアーキテクチャ

第4章

関連研究

第5章

GoogLeNetの並列化

第6章

実装

6.1 実装環境

6.2 実装方針

6.3 並列化戦略

第7章

評価

7.1 評価環境

7.2 性能評価

7.3 電力効率評価

第8章

まとめと今後の課題

8.1 結論

マルチ FPGA システムに GoogLeNet を実装し評価をした．その結果，性能比で対 CPU ○○倍, 対 GPU ○○倍を実現し，電力効率でも対 CPU ○○倍, 対 GPU ○○倍を実現した．

8.2 今後の課題

本論文では基本的な高速化の手法しか提案しなかった．本研究を踏まえ，マルチ FPGA のネットワークの形やアクセラレータのさらなる研究，そして各アクセラレータのロードバランスの検討を行い，より高速化，電力効率向上を図る必要がある

第9章

謝辞

本研究に取り組むにあたりご指導ご鞭撻を賜りました天野英晴教授，に深く感謝いたします.