

# CbCを用いたPerl6処理系

清水 隆博<sup>1,a)</sup> 河野 真治<sup>1,b)</sup>

**概要:** スクリプト言語であるPerl5の後継言語としてPerl6が現在開発されている。Perl6は設計と実装が区分されており様々な処理系が開発されている。現在主流なPerl6はRakudoと言われるプロジェクトである。RakudoではPerl6自体をNQP(NotQuitePerl)と言われるPerl6のサブセットで記述し、NQPをVMが解釈するという処理流れになっている。このVMは任意のVMが選択できるようになっており、現在はMoarVM,JavaVM,JavaScriptが動作環境として選択可能である。主に利用されているVMにCで書かれたMoarVMが存在する。MoarVMはJITコンパイルなどをサポートしているが、全体的な起動時間及び処理速度がPerl5と比較し非常に低速である。この問題を解決するためにContinuation based C (CbC) という言語を一部用いてMoarVMの書き換えを行う。本論文ではCbCを用いたMoarVMの書き換えを検討し、得られた知見について述べる。

**キーワード:** プログラミング言語, コンパイラ, CbC, Perl6, MoarVM

## 1. CbCを用いた言語処理系の実装

当研究室ではContinuation Based C(以下CbC)という言語を開発している。CbCはCよりきめ細やかな単位で実装する事が可能である為、言語処理系に応用すれば効率的な開発、実行が出来ると期待される。現在活発に開発が進んでいる言語にPerl6がある。Perl6はMoarVMと呼ばれるVMを中心としたRakudoと呼ばれる実装が現在の主流となっている。Rakudoは処理速度が他のプログラミング言語と比較しても非常に低速である。その為現在日本国内ではPerl6を実務として利用するケースは概ね存在しない。Perl6の持つ言語機能や型システムは非常に柔軟かつ強力であるため実用的な処理速度に達すれば言語の利用件数が向上することが期待される。その為本研究では、CbC

を用いた言語処理系の実装の一例としてMoarVMをCbCで構築したCbCMoarVMを実装する。本研究はCbCをスクリプト言語の実装に適用した場合、どのような利点やプログラミング上の問題点に遭遇するかCbCの応用としての側面でも行う。本稿ではまずCbC, Perl6の特徴及び現在の実装について述べ、本研究で行ったCbCで書き換えたMoarVMについてデバッグ手法も含め解説する。そして本研究で得られたCbCを言語処理系に適用した場合の利点と欠点について述べ、今後の展望について記載する。

## 2. CbC

### 2.1 CbCの概要

CbCは当研究室で開発しているプログラミング言語である。Cレベルでのプログラミングを行う場合、本来プログラマが行いたい処理の他にmallocなどを利用したメモリのアロケートやエラー

<sup>1</sup> 琉球大学工学部情報工学科

a) anatofuz@cr.ie.u-ryukyu.ac.jp

b) kono@ie.u-ryukyu.ac.jp

ハンドリングなどを記述する必要がある。これらの処理を meta computation と呼ぶ。これら meta computation と通常の処理を分離することでバグの原因が meta computation 側にあるか処理側にあるかの分離などが可能となる。しかし C 言語などを用いたプログラミングで meta computation の分離を行おうとすると、それぞれ事細かに関数やクラスを分割せねばならず容易ではない。CbC では関数より meta computation を細かく記述する為に CodeGear という単位を導入した。また CodeGear の実行に必要なデータを DataGear という単位で受け渡す。CbC では CodeGear, DataGear を基本単位として記述するプログラミングスタイルを取る。

## 2.2 CodeGear と DataGear

CbC では C の関数の代わりに CodeGear を導入する。CodeGear は C の関数宣言の型名の代わりに `__code` と書くことで宣言できる。`__code` は CbC コンパイラの扱いは `void` と同じ型であるが、CbC プログラミングでは CodeGear であることを示す識別子としての意味で利用する。CodeGear 間の移動は `goto` 文によって記述する。

```
extern int printf(const char*,...);

int main(){
    int data = 0;
    goto cg1(&data);
}

__code cg1(int *datap){
    (*datap)++;
    goto cg2(datap);
}

__code cg2(int *datap){
    (*datap)++;
    printf("%d\n",*datap);
}
```

Code 1: cbc\_example.cbc

Code??に示す CbC のコードでは main 関数から `cs1`, `cs2` に遷移し、最終的に `data` の値が 2 となる。CodeGear 間の入出力の受け渡しは引数を利用する。この引数は小さな DataGear であると言える。

## 2.3 軽量継続

CbC では次の CodeGear に移行する際、C の `goto` 文を利用する。通常の C の関数呼び出しの場合、スタックポインタを操作しローカル変数などをスタックに保存する。CbC の場合スタックフレームを操作せず、レジスタの値を変更せずそのまま次の CodeGear に遷移する事が可能である。通常 Scheme の `call/cc` などの継続は現在の位置までの情報を環境として所持した状態で遷移する。対して CbC は環境を持たず遷移する為、通常の継続と比較して軽量であることから軽量継続であると言える。CbC は軽量継続を利用するためレジスタレベルでのきめ細やかな実装が可能となっている。

## 2.4 現在の実装

CbC は現在主要な C コンパイラである `gcc` 及び `llvm` をバックエンドとした `clang` 上の 2 種類の実装が存在する。`gcc` はバージョン 9.0.0 に、`clang` は 7.0.0 に対応している。

## 2.5 CbC コンパイラのバグ

CbC コンパイラは現在も開発中であり幾つかのバグが発見されている。まず CodeGear 内で宣言した局所変数のポインタを別の変数などで確保した状態で `goto` してしまうと tail call 最適化が切られる。これはただの関数呼び出しになってしまう為、直接的な被害はないものの CbC としての利点が損なわれてしまう。また本来は操作しないはずのスタック領域の操作が入ってしまうため、プログラマの意図と反したスタックポインタなどの操作が行われてしまいバグが発生する可能性が存在する。

## 2.6 CbC と C の互換性

CbC コンパイラは内部的に与えられているソースコードが CbC であるかどうかを判断する。この際に CodeGear を利用していない場合は通常の C プログラムとして動作する。その為今回検証する MoarVM のビルドにおいても CbC で書き換えたソースコードがある MoarVM と、手を加えていないオリジナルの MoarVM の 2 種類を同一の CbC

コンパイラでビルドする事が可能である。

また C から CbC への遷移時に、再び C の関数に戻るように実装したい場合がある。その際は環境付き goto と呼ばれる手法を取る。これは `_CbC_return` 及び `_CbC_environment` という変数を渡す。この変数は `_CbC_return` が元の環境に戻る際に利用する `CodeGear` を指し、`_CbC_environment` は復帰時に戻す元の環境である。復帰する場合、呼び出した位置には帰らず、呼び出した関数の終了する位置に戻る。

```
__code cg(__code (*ret)(int,void *),void *env
    ){
    goto ret(1,env);
}

int c_func(){
    goto cg(_CbC_return,_CbC_environment);
    return -1;
}

int main(){
    int test;
    test = c_func();
    printf("%d\n",test);
    return 0;
}
```

Code 2: 環境付き継続の例

Code??に示す例では `c_func` から環境付き継続で `cs` に継続している。通常 `c_func` の戻り値は `-1` であるが、`cs` から環境付き継続で `main` に帰る為に `cs` から渡される `1` が `test` の値となる。

## 2.7 言語処理系における CbC の応用

CbC を言語処理系、特にスクリプト言語に応用すると幾つかの箇所に置いて利点があると推測される。CbC における `CodeGear` はコンパイラの基本ブロックに相当する。その為従来のスクリプト言語では主に `case` 文で記述していた命令コードディスパッチの箇所を `CodeGear` の遷移として記述する事が可能である。CbC は状態を単位として記述が可能であるため、命令コードなどにおける状態を利用するスクリプト言語の実装は応用例として適していると考えられる。

## 3. Perl6 の概要

この章では現在までの Perl6 の遍歴及び Perl6 の言語的な特徴について記載する。

### 3.1 Perl6 の構想と初期の処理系

Perl6 は 2002 年に LarryWall が Perl を置き換える言語として設計を開始した。Perl5 の言語的な問題点であるオブジェクト指向機能の強力なサポートなどを取り入れた言語として設計された。Perl5 は設計と実装が同一であり、Larry らによって書かれた C 実装のみだった。Perl6 は設計と実装が分離しており様々な処理系が開発されてきた。まず 2005 年に唐鳳によって Haskell で実装された Pugs?が登場した。Pugs は最初に登場した Perl6 実装であり、この実装を基にして Perl6 の仕様も修正された。現在 Pugs は歴史的な実装となっており、更新はされていない。

### 3.2 Parrot

その後 Python との共同動作環境として Parrot?が実装された。Parrot は PASM と呼ばれるバイトコードを解釈可能なレジスタマシンである。Parrot での Perl6 の実装は NQP(NotQuitPerl) と呼ばれる Perl6 のサブセットで Perl6 を記述するというアイデアの基実装された。ParrotVM は 2006 年の version8.1.0 が最後のリリースである。こちらも Pugs と同様に現在の Perl6 プロジェクトでは歴史的な実装とされている。現在主に使用されている実装である Rakudo は 2010 年に Rakudo-Star という一連のツール郡としてリリースされた。

Perl6 は言語仕様及び処理実装が Perl5 と大幅に異なっており、言語的な互換性が存在しない。従って現在では Perl6 と Perl5 は別言語としての開発方針になっている。Perl6 は現在有力な処理系である Rakudo から名前を取り Raku という別名がつけられている。

### 3.3 Rakudo

Rakudo とは Parrot で構想に上がった NQP,

NQP に基づく Perl6 を基にしたプロジェクトである。Rakudo が Perl6 のコンパイラかつインタプリタであると考えても良い。Rakudo は図??に示す構成になっている。Rakudo におけるコンパイラとは厳密には 2 種類存在する。まず第 1 のものが Perl6、もしくは NQP を MoarVM、JVM のバイトコードに変換する NQP コンパイラである。次にその NQP が出力したバイトコードをネイティブコードに変換する VM の 2 種類である。この VM は現在 MoarVM、JavaVM、Javascript を選択可能である。Rakudo 及び NQP project ではこの NQP コンパイラの部分をフロントエンド、VM の部分をバックエンドと呼称している。NQP で主に書かれ、MoarVM など NQP が動作する環境で動く Perl6 のことを Rakudo と呼ぶ。Perl6 は NQP 以外にも NQP を拡張した Perl6 自身で書かれている箇所が存在し、これは NQP コンパイラ側で MoarVM が解釈可能な形へ変換を行う。

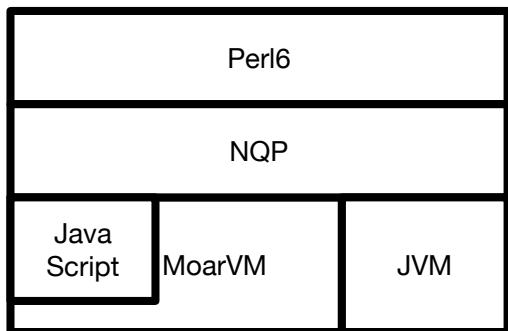


図 1: Perl6 の構成

### 3.4 NQP

Rakudo における NQP は現在 MoarVM、JVM 上で動作し、MoarVM を一部利用することで NodeJS から動作させる事が可能である。NQP は Perl6 のサブセットであるため主な文法などは Perl6 に準拠しているが幾つか異なる点が存在する。NQP は最終的には NQP 自身でブートストラップする言語であるが、ビルドの最初にはすでに書かれた MoarVMByteCode を必要とする。この MoarVMByteCode の状態を Stage0 と言い、ディ

レクトリ名として設定されている。Perl6 の一部は NQP を拡張したもので書かれている為、Rakudo を動作させる為には MoarVM などの VM、VM に対応させる様にビルドした NQP がそれぞれ必要となる。現在の NQP では MoarVM、JVM に対応する Stage0 はそれぞれ MoarVMBytecode.jar ファイルが用意されており、Javascript ではバイトコードの代わりにランタイム独自の ModuleLoader などが設計されている。MoarVM の ModuleLoader は Stage0 にある MoarVMBytecode で書かれた一連のファイルが該当する。

Stage0 にあるファイルを MoarVM に与えることで nqp のインタプリタが実行される様になっている。これは Stage0 の一連のファイルは MoarVMBytecode など記述された NQP コンパイラのモジュールである為である。NQP は 6model と呼ばれるオブジェクトモデルを採用しているが、これを構築する為に必要な NQPCORE、正規表現系の QRegex、MoarVM の ModuleLoader などが MoarVMBytecode で記述されている。これら MoarVMBytecode の拡張子は MoarVM である。MoarVM に対して Stage0 のディレクトリにライブラリパスを設定し、nqp.MoarVM を実行させることで nqp の対話型環境が起動する。

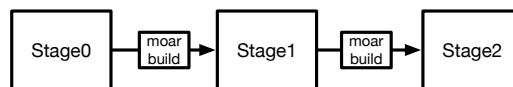


図 2: NQP のビルドフロー

NQP のビルドフローを図??に示す。実際に perl6 を動かすためには self build した NQP コンパイラが必要となる。その為に stage0 を利用して Stage1 をビルドし NQP コンパイラを作成する。Stage1 は中間的な出力であり、生成された NQP ファイルは Stage2 と同一であるが、MoarVMBytecode が異なる。Perl6 では完全なセルフコンパイルを実行した NQP が要求される為、Stage1 を利用してもう一度ビルドを行い Stage2 を作成する。

Perl6 のテストスイートである「Roast?」やドキュメントなどによって設計が定まっている Perl6

とは異なり NQP 自身の設計は今後も変更になる可能性が開発者から公表されている。現在の公表されている NQP のオペコードは NQP のリポジトリ?に記述されているものである。

### 3.5 Rakudo Perl6

Rakudo 実装上における Perl6 は Rakudo Perl6 と呼ばれている Git リポジトリで管理されているプログラムのことである。前述した通り Rakudo Perl6 は Perl6 のサブセットである NQP を用いて記述されている。従って yacc や lex と呼んだ Perl5 の文字解析、構文解析に利用していたプログラムは利用せず、NQP 側で構文定義などを行っている。NQP は NQP 自身で Bootstrapping されている為、Rakudo Perl6 の build 時には NQP の実行環境として要した VM、それに基づいて build した NQP がそれぞれ必要となる。

言語的な特徴としては独自に Perl6 の文法を拡張可能な Grammar, Perl5 と比較した場合のオブジェクト指向言語としての進化も見られる。また Perl6 は漸進的型付け言語である。従来の Perl の様に変数に代入する対象の型や文脈に応じて型を変更する動的型言語としての側面を持ちつつ独自に定義した型を始めとする様々な型に静的に変数の型を設定する事が可能である。

### 3.6 現在の Perl6

Perl6 の言語仕様?とその時点での実装状況をまとめた公式ドキュメント?は分離している。従来は言語仕様は自然言語の仕様書であったが、現在はテストスイートである「Roast?」そのものと定義されている。現在の Perl6 の仕様を読む場合 Roast を確認し、現在 rakudo 上に実装されている機能を見る場合は公式ドキュメントを確認する必要がある。

### 3.7 処理速度

現在の Perl6 が他のプログラミング言語と比較した場合どのような違いがでるのか計測した。mac os の /var/log/system.log ファイルから正規表現でログ中のプログラムが書き込んだ回数を個別に数え上げるというものである。今回はファイルを

231K と 3GB の二種類用意し、どのような違いが出るのか測定した。

測定した環境は次の通りである。

- Perl6 (MoarVM) ver.2018.04.01
- Perl6 (JVM) 2018.06-163-g612d071b8 built on JVM
- Python 3.6.5
- Java 10
- Perl5

測定した結果を以下に示す。測定結果の単位は秒である。

| FileSize | MoarVM  | Perl6 on JVM | Python3 | Java  | Perl5 |
|----------|---------|--------------|---------|-------|-------|
| 231K     | 0.86    | 21.48        | 0.06    | 0.27  | 0.04  |
| 3G       | 2331.08 | 1665.56      | 101.16  | 48.85 | 41.35 |

計測結果からファイルサイズが小さい場合は MoarVM より JVM に乗せた Perl6 が低速であるが、ファイルサイズが大きい場合は Java の JIT が働くため MoarVM より高速に動いていると推測できる。

## 4. CbC による MoarVM

この章では改良を行った Perl6 処理系である MoarVM について述べる。今回改良を行った MoarVM は 2018.04.01, nqp は 2018.04-3-g45ab6e3 バージョンで行っている。

### 4.1 方針

MoarVM そのものを CbC で書き換えることも考えられるが MoarVM 自体既に巨大なプロジェクトである為すべてを書き換える事は困難である。その為まず比較的 CbC で書き換えることが容易な箇所を修正する。前章までに述べた通り CbC の CodeGear はコンパイラの基本ブロックに該当する。従って MoarVM における基本ブロックの箇所を CodeGear に書き換える事が可能である。MoarVM における基本ブロックはインタプリタが実行するバイトコードごとの処理に該当する。

## 4.2 MoarByteCode のディスパッチ

MoarVM のバイトコードインタプリタは `src/core/interp.c` で定義されている。この中の関数 `MVM_interp_run` で命令に応じた処理を実行する。関数内では命令列が保存されている `cur_op`、現在と次の命令を指し示す `op`、Thread の環境が保存されている `Threadcontext` などの変数を利用する。命令実行は大きく二種類の動作があり、C の `goto` が利用できる場合は `Code??` に示す `MVM_CGOTO` フラグが立ちラベル遷移を利用する。それ以外の場合は巨大な `case` 文として命令を実行する。

ラベル遷移を利用する場合は `Code??` に示すラベルテーブル `LABELS` にアクセスし、テーブルに登録されているアドレスを取得し、マクロ `NEXT` で遷移する。`Code??` に示す `no_op` は何もせず次の命令に移動する為、`NEXT` のみ記述されている。

このラベルテーブルの中身はラベルが変換されたアドレスであるため、実際に呼ばれている命令コードの名前はデバッガレベルでは確認できない。C レベルでのデバッグ時にはアドレスと実際に呼ばれる箇所を確認する事に手間がかかる。巨大な `case` 文として実行された場合、実行時間が遅いだけでなく、ラベル遷移と共存させて記述を行っている為 C のソースコードにおける可読性も低下する。

```
static const void * const LABELS[] = {
    &&OP_no_op,
    &&OP_const_i8,
    &&OP_const_i16,
    &&OP_const_i32,
    &&OP_const_i64,
    &&OP_const_n32,
    &&OP_const_n64,
    &&OP_const_s,
    &&OP_set,
    &&OP_extend_u8,
    &&OP_extend_u16,
    &&OP_extend_u32,
    &&OP_extend_i8,
    &&OP_extend_i16,
```

Code 3: ラベルテーブルの一部

```
#define NEXT_OP (op = *(MVMuint16 *) (cur_op),
               cur_op += 2, op)

#if MVM_CGOTO
```

```
#define DISPATCH(op)
#define OP(name) OP_ ## name
#define NEXT *LABELS[NEXT_OP]
#else
#define DISPATCH(op) switch (op)
#define OP(name) case MVM_OP_ ## name
#define NEXT runloop
#endif
```

Code 4: `interp.c` のマクロ部分

```
DISPATCH(NEXT_OP) {
    OP(no_op):
        goto NEXT;
    OP(const_i8):
    OP(const_i16):
    OP(const_i32):
        MVM_exception_throw_adhoc(tc, "
            const_iX_MYI");
    OP(const_i64):
        GET_REG(cur_op, 0).i64 =
            MVM_BC_get_I64(cur_op, 2);
        cur_op += 10;
        goto NEXT;
    OP(pushcompsc): {
        MVMObject * const sc = GET_REG(
            cur_op, 0).o;
        if (REPR(sc)->ID !=
            MVM_REPR_ID_SCREf)
            MVM_exception_throw_adhoc(
                tc, "Can only push an
                SCREf with pushcompsc");
        ;
        if (MVM_is_null(tc, tc->
            compiling_scs)) {
            MVMROOT(tc, sc, {
                tc->compiling_scs =
                    MVM_repr_alloc_init
                        (tc, tc->instance->
                            boot_types.
                                BOOTArray);
            });
        }
        MVM_repr_unshift_o(tc, tc->
            compiling_scs, sc);
        cur_op += 2;
        goto NEXT;
    }
}
```

Code 5: オリジナル版 MoarVM のバイトコードディスパッチ

`interp.c` では命令コードのディスパッチはマクロを利用した `cur_op` の計算及びラベルの遷移、も

しくはマクロ DISPATCH が展開する switch 文で行われていた。CbCMoarVM ではこの問題を解決するために、それぞれの命令に対応する CodeGear を作成し、CodeGear 名前を要素として持つ CbC の CodeGear のテーブルを作成した。この CodeGear のテーブルを参照する CodeGear は cbc\_next であり、この中のマクロ NEXT は interp.c のマクロ NEXT を CbC 用に書き直したものである。

```
#define NEXT_OP(i) (i->op = *(MVMuint16 *) (i
->cur_op), i->cur_op += 2, i->op)

#define DISPATCH(op) {goto (CODES[op])(i);}
#define OP(name) OP_ ## name
#define NEXT(i) CODES[NEXT_OP(i)](i)
static int tracing_enabled = 0;

__code cbc_next(INTERP i){
    goto NEXT(i);
}
```

Code 6: CbCMoarVM のバイトコードディスパッチ

### 4.3 命令実行箇所の CodeGear への変換

ラベルテーブルや case 文の switch 相当の命令実行箇所を CbC に変換し、CodeGear の遷移として利用する。interp.c は Code?? に示すスタイルで記述されている。

OP(\*) の \* に該当する箇所はバイトコードの名前である。通常このブロックには LABEL から遷移する為、バイトコードの名前は LABELS の配列の添字に変換されている。そのため対象となる CodeGear を LABELS の並びと対応させ、Code?? に示す CodeGear の配列 CODES として設定すれば CodeGear の名前は問わない。今回は CodeGear である事を示す為に suffix として cbc\_をつける。

```
__code (* CODES[])(INTERP) = {
    cbc_no_op,
    cbc_const_i8,
    cbc_const_i16,
    cbc_const_i32,
    cbc_const_i64,
    cbc_const_n32,
    cbc_const_n64,
    cbc_const_s,
    cbc_set,
```

```
    cbc_extend_u8,
    cbc_extend_u16,
```

Code 7: CodeGear 配列の一部

命令の実行処理で MoarVM のレジスタである reg\_base や命令列 cur\_op などの情報を利用しているが、これらは MVM\_interp\_run 内のローカル変数として利用している。ラベルを利用しているオリジナル版では同一関数内であるためアクセス可能であるが、CodeGear 間の移動で命令を表現する CbC ではアクセスできない。その為インタプリタの情報を集約した構造体 inter を定義し、この構造体へのポインタである INTERP 型の変数 i を CodeGear の入出力として与える。CodeGear 内では INTERP を経由することでインタプリタの各種情報にアクセスする。CodeGear 間の遷移ではレジスタの値の調整は行われなない為、入力引数を使ってレジスタマッピングを管理できる。その為 INTERP のメンバである MoarVM のレジスタそのものをアーキテクチャのレジスタ上に乗せる事が可能である。

命令実行中の CodeGear の遷移を図?? に示す。この中で実線で書かれている部分は CbC の goto 文で遷移し、波線の箇所は通常の C の関数呼び出しとなっている。

現在の CbCMoarVM は次の命令セットのディスパッチを cbc\_next が行っていた。その為 cbc\_next から命令コードに対応する CodeGear に継続し、CodeGear から cbc\_next に継続するサイクルが基本の流れである。CodeGear 内から C の関数は問題なく呼ぶ事が可能であるため、C の関数を利用する処理は変更せず記述する事ができる。また変換対象は switch 文であるため、break せず次の case に移行した場合に対応するように別の CodeGear に継続し、その後 cbc\_next に継続するパターンも存在する。

```
__code cbc_no_op(INTERP i){
    goto cbc_next(i);
}
__code cbc_const_i8(INTERP i){
    goto cbc_const_i16(i);
}
```

```

__code cbc_const_i16(INTERP i){
    goto cbc_const_i32(i);
}
__code cbc_const_i32(INTERP i){
    MVM_exception_throw_adhoc(i->tc, "const_iX
    NYI");
    goto cbc_const_i64(i);
}
__code cbc_const_i64(INTERP i){
    GET_REG(i->cur_op, 0,i).i64 =
        MVM_BC_get_I64(i->cur_op, 2);
    i->cur_op += 10;
    goto cbc_next(i);
}
__code cbc_pushcompssc(INTERP i){
    static MVMObject * sc;
    sc = GET_REG(i->cur_op, 0,i).o;
    if (REPR(sc)->ID != MVM_REPR_ID_SCSRef)
        MVM_exception_throw_adhoc(i->tc, "Can
        only push an SCSRef with pushcompssc
        ");
    if (MVM_is_null(i->tc, i->tc->
        compiling_scs)) {
        MVMROOT(i->tc, sc, {
            i->tc->compiling_scs =
                MVM_repr_alloc_init(i->tc, i->
                tc->instance->boot_types.
                BOOTArray);
        });
    }
    MVM_repr_unshift_o(i->tc, i->tc->
        compiling_scs, sc);
    i->cur_op += 2;
    goto cbc_next(i);
}

```

Code 8: CbCMoarVM のバイトコードに対応する CodeGear

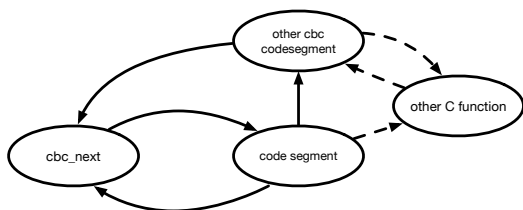


図 3: CbC における MoarVM バイトコードインタプリタ内の状態遷移

バイトコードの数は膨大である為、すべてを手作業で変換する事は望ましくない。本研究では PerlScript を用いて interp.c から CbC の

CodeGear を自動生成するスクリプトを作成した。このスクリプトでは以下の修正手続きを実行する。

- OP(\*) の.\*部分を CodeGear の名前として、先頭に cbc\_をつけた上で設定する。
- cur\_op など構造体 INTER のメンバ変数はポインタ i から参照するように修正する
- GC 対策のためマクロ MVMROOT を使い局所変数のポインタをスタックに積む箇所は、局所変数を static 化する
- 末尾の goto NEXT を goto cbc\_next(i) に修正する
- case 文で下の case 文に落ちている箇所は、case 文に対応する CodeGear に遷移する様に goto 文を付け加える

上記 Code??では cbc\_const\_i8 などが case 文の下の case 部分に該当する cbc\_const\_i64 に遷移する様に変換されている。また cbc\_pushcompssc では MVMROOT に局所変数 sc を渡している為、これを static で宣言し直している。

現在 CbC で記述された OS である GearsOS には Interface が導入されている。これは Java の interface, Haskell の型クラスに該当する概念であり、次の CodeGear に Interface 経由で継続する事が可能である。Interface は現在の MoarVM には実装されていない為、今後 ThreadeCode の実装を行うにあたり導入を検討している。

## 5. MoarVM のデバッグ

MoarVM 自体のデバッグは MoarVM のリポジトリにテストコードが付随していない為単体では実行不可能である。本研究では MoarVM のデバッグにおける C デバッガの使用方法和 MoarVM のテスト方法についても考案する。

### 5.1 MoarVM の Bytecode のデバッグ

moar に対して MoarVM bytecode を dump オプションを付けて読み込ませると MoarVM の bytecode がアセンブラの様に出力される。しかしこれは MoarVM が実行した bytecode のトレースではなく MoarVM bytecode を変換したものに過ぎな



い。また、明らかに異なる挙動を示す両者の moar を利用しても同じ結果が返ってきてしまう。そのため今回の MoarVMBytecode インタプリタの実装のデバッグにはこの方法は適さない。従って実際に実行した命令を確認するには gdb などの C デバッガを利用して MoarVM を直接トレースする必要がある。

CbC 側は Code?? に示す様に cbc\_next に break point を設定する。オリジナル側は次のオペコードの設定のマクロにダミーの関数を呼び出すように修正し、そこに break point を設定する。CbC 側では CodeGear の名前をデバッグ上で直接確認できるが、オリジナル版は LABEL の配列の添え字から自分でどのオペコードに対応しているかをデバッガの外で探さなければならない。

添字を確認するためには Code?? に示すようにオリジナルの MoarVM の場合 cur\_op の値を MV-Muint16 のポインタでキャストし、これが指す値を出力する。break point を掛けているダミー関数では cur\_op にアクセスする事が出来ない為、スタックフレームを一つ up する必要がある。

```
(gdb) b cbc_next
Breakpoint 2 at 0x7ffff7560288: file src/core/cbc-interp.cbc, line 61.
(gdb) command 2
Type commands for breakpoint(s) 2, one per line.
End with a line saying just "end".
>p CODES[(MVMuint16 *)i->cur_op]
>p *(MVMuint16 *)i->cur_op
>c
>end
```

Code 9: CbCMoarVM に対しての break point 設定

```
dalmore gdb --args ../../MoarVM_Original/MoarVM/moar --libpath=src/vm/moar/stage0/gen/moar/stage1/nqp
(gdb) b dummy
Function "dummy" not defined.
Make breakpoint pending on future shared library load? (y or [n]) y
Breakpoint 1 (dummy) pending.
(gdb) command 1
Type commands for breakpoint(s) 1, one per line.
End with a line saying just "end".
```

```
>up
>p *(MVMuint16 *) (cur_op)
>c
>end
```

Code 10: オリジナル版 MoarVM に対しての break point 設定

## 5.2 MoarVM の並列デバッグ手法

しかし MoarVM が実行する命令は膨大な数がある。その為 gdb で MoarVM を CbC とオリジナル版での並列デバッグを人間が同時に行うことは困難である。Perl などのスクリプトを用いて自動的に解析したいため、ログを残す為に script コマンドを実行した状態で gdb を起動する。トレースでは実行した命令名のみ取得できれば良い為、Code??, ?? で debug point に command として設定している様に、設定された cur\_op の値を出力し続けるのみの動きを導入する。

実際に実行したログ・ファイルの一部をそれぞれ Code??, ?? に示す。

```
Breakpoint 1, dummy () at src/core/interp.c:46
46 }
#1 0x00007ffff75608fe in MVM_interp_run (tc=0x604a20,
initial_invoke=0x7ffff76c7168 <
toplevel_initial_invoke>, invoke_data=0x67ff10)
at src/core/interp.c:119
119 goto NEXT;
$1 = 159

Breakpoint 1, dummy () at src/core/interp.c:46
46 }
#1 0x00007ffff75689da in MVM_interp_run (tc=0x604a20,
initial_invoke=0x7ffff76c7168 <
toplevel_initial_invoke>, invoke_data=0x67ff10)
at src/core/interp.c:1169
1169 goto NEXT;
$2 = 162
```

Code 11: オリジナル版 MoarVM のバイトコードのトレース

```

Breakpoint 2, cbc_next (i=0x7fffffffdc30) at
  src/core/cbc-interp.cbc:61
61 goto NEXT(i);
$1 = (void (*)(INTERP)) 0x7ffff7566f53 <
  cbc_takeclosure>
$2 = 162

Breakpoint 2, cbc_next (i=0x7fffffffdc30) at
  src/core/cbc-interp.cbc:61
61 goto NEXT(i);
$3 = (void (*)(INTERP)) 0x7ffff7565f86 <
  cbc_checkarity>
$4 = 140

Breakpoint 2, cbc_next (i=0x7fffffffdc30) at
  src/core/cbc-interp.cbc:61
61 goto NEXT(i);
$5 = (void (*)(INTERP)) 0x7ffff7579d06 <
  cbc_paramnamesused>
$6 = 558

```

Code 12: CbCMoarVM のバイトコードのトレース

オリジナル版では実際に実行する命令処理はレベルに変換されてしまう為名前をデバッガ上では出力できないが、CbC では出力する事が可能である。CbC とオリジナルの CODES, LABEL の添字は対応している為、ログの解析を行う際はそれぞれの添字を抽出し違いが発生している箇所を探索する。これらは script コマンドが作成したログを元に異なる箇所を発見するスクリプトを用意し自動化する。(Code ??)

```

131 : 131
139 : 139
140 : 140
144 : 144
558 : 558
391 : 391
749 : 749
53 : 53
*54 : 8

```

Code 13: バイトコードの差分検知の一部分

違いが生じている箇所が発見できた場合、その前後の CodeGear 及びディスパッチ部分に breakpoint をかけ、それぞれの変数の挙動を比較する。主に cbc.return 系の命令が実行されている場合は、その直前で命令を切り替える cbc.invoke 系統の命令が呼ばれているが、この周辺で何かしらの

違いが発生している可能性が高い。また主に次の CodeGear に遷移する際に CbC コンパイラのバグが生じている可能性もある為、アセンブラレベルの命令を確認しながらデバッグを進めることとなる。

### 5.3 MoarVM のテスト方法

MoarVM は単体で実行する事が不可能である。また NQP のリポジトリに付随するテストは nqp で書かれている為、MoarVM のテストは NQP リポジトリのファイルを直接使う場合は出来ない。今回は正常に動く MoarVM, NQP を用意し nqp を経由してデバッグ

### 5.4 CbC コンパイラによるバグ

現在までの CbC は複数個の入出力を CodeGear に与えるユースケースで利用していた。CbC コンパイラ自身はそれぞれ用意したテストスイートを通化するものの、MoarVM の様な巨大なプロジェクトの CS をコンパイルを実行する場合、予期せぬバグが発生した。主に CodeGear 間の goto における tail call フラグの除去や、DataGear として渡している構造体の変数のアドレスがスタックポインタの値より上位に来てしまい、通常の C の関数を call した際にローカル変数の領域が DataGear のアドレスの周辺を利用してしまふ。その為 DataGear の構造体の値が書き換わり、C から DataGear に return した際に DataGear の構造体が破壊されるバグである。このバグは先程の並列デバッグを行いながらプログラムカウンタや変数の動きをトレースする事などで発見することが出来る。現状では CbC コンパイラがプログラムの意図と反する挙動を取るため CbC コンパイラのバグを回避するプログラミングが要求されている。本来コンパイラ側のバグを回避するプログラミングをプログラマに要求する事は好ましくない。従って CbC コンパイラ自身の信頼性を向上させる事も今後の課題となっている。

また現在は clang 上に実装した CbC コンパイラでは CodeGear 内部の tail call 除去のエラーが発生してしまう為コンパイルする事が出来ない。その為現在は gcc 上に実装した cbc コンパイラを利

用し gdb を利用しデバッグを行う。

## 6. CbCMoarVM の利点と欠点

MoarVM の様な巨大なスクリプト言語処理系に CbC を適応した所現在までに複数の利点と欠点が発見された。本章ではまず利点を述べ、次に現段階での CbC を適応した場合の欠点について考察する。

オリジナルの MoarVM では命令コードに対応する箇所はラベルジャンプ、もしくは switch 文で実装されていた。その為同じ C ファイルに命令コードの実行の定義が存在しなければならない。今後 MoarVM に新たなバイトコードが導入されていく事を考えると interp.c が巨大になる可能性がある。関数単位での処理の比重が偏る事に加え、switch 文中に書かれている処理は他の関数から呼ぶ事が出来ないため、余計な手間がかかる箇所が発生すると考えられる。

CbCMoarVM の場合、CodeGear として基本ブロックを記述出来る為オリジナルの MoarVM の様に switch 文のブロック中に書く必要性が無くなる。その為類似する命令系をコード分割し、モジュール化する事が可能である。また CbC は goto 文で遷移する以外は通常の C の関数と同じ扱いをする事が可能である。従って CodeGear 内部の処理を別の箇所から使用する事も可能となる為再利用性も向上する。

ThrededCode を実装する場合、通常命令ディスパッチの箇所と、実際に実行される命令処理を大幅に変更しなければならない。CbC を用いた実装の場合、命令処理はただの CodeGear の集合である。その為 CodeGear を ThrededCode に対応した並びとして選択する事ができれば命令処理部分の修正をほぼせずに ThrededCode を実現する事が可能である。

また CodeGear はバイトコードレベルと同じ扱いができるため、ThrededCode そのものを分離して最適化をかける事が可能である。これも CodeGear が関数単位として分離できる事からの利点である。

MoarVM のバイトコードインタプリタの箇所はオリジナルの実装ではラベルジャンプを用いて実

装されている。その為、直接ラベルに break point をかける事が出来ない。作業者がデバッガが読み込んでいる C ソースコードの位置を把握し、行番号を指定して debug point を設定する必要があった。

CbCMoarVM の場合、CodeGear 単位でバイトコードの処理単位を記述している為、通常関数と同じく直接 CodeGear にデバッグポイントをかける事が可能である。これは C プログラミングの関数に対してのデバッグで、状態ごとに break point をかける事が出来ることを意味する。通常の C 言語で言語処理系を実装した場合と比較して扱いやすくなっていると言える。さらにラベルテーブルでの管理場合、次のバイトコード箇所は数値でしか確認できず、実際にどこに飛ぶのかはラベルテーブル内と数値を作業者が手作業で確認する必要があった。スクリプトなどを組めば効率化は出来るがデバッガ上で完結しない為手間がかかる。CbC 実装では CODES テーブル内は次の CodeGear の名前が入っている為、数値から CodeGear の名前をデバッガ上で確認する事が出来る。

現在 MoarVM は LuaJit? を搭載し JIT コンパイラを行っている。LuaJIT そのものを CbC に適応させるわけではないが、CbC の ABI に JIT されたコードを合わせる事が可能であると推測できる。

本来処理系は広く使われる為に著名な OSS などを利用して開発するのが良いが、CbC プロジェクトの認知度が低いという現状がある。

また、前章までに複数述べた通り CbC コンパイラが現在非常にバグを発生させやすい状態になっている。CbC コンパイラは gcc と llvm/clang に実装している為、これらのアップデートに追従する必要がある。しかしコンパイラのバージョンに応じて CbC で利用するコンパイラ内の API が異なる場合が多く、API の変更に伴う修正作業などを行う必要がある。

CbCMoarVM では C から CodeGear へ、CodeGear から C への遷移などが複数回繰り返されているが、この処理中の CodeGear での tail call の強制が非常に難関である。tail call の強制には関数定義の箇所や引数、スタック領域のサイズ修正などを行う必要がある。現在のバグでは

CodeGear 内部での不要なスタック操作命令を完全に排除しきれていない。

また CodeGear から C に帰る場合、環境付き継続を行う必要がある。C の関数の末尾で CodeGear を呼び出している場合など環境付き継続を使用しなくても良いケースは存在するが、頻繁に C と CbC を行き来する場合記述が冗長になる可能性はある。

## 7. Threaded Code

現在の MoarVM は次の命令をバイトコードからディスパッチし決定後、ラベルジャンプを利用して実行している。この処理ではディスパッチの箇所にコストが掛かってしまう。CbC を MoarVM に導入することで、バイトコード列を直接サブルーチンコールの列に置き換えてしまう事が可能である。これは CbC が基本ブロックの単位と対応している為である。CbC では現在ディスパッチを行う CodeGear である `cbc_next` を利用しているが、Threaded Code を実装するにあたり、`cbc_next` と次の CodeGear に直接遷移する `cbc_fixt_next` の実装を予定している。

また段階的に現在 8 バイト列を 1 命令コードとして使用しているが、これを 16 バイトなどに拡張し 2 命令を同時に扱えるように実装する事なども検討している。

Perl5 においては `perlcc` というモジュールが開発されている。これは Perl5 内部で利用している Perl バイトコードを、Perl の C API である XS 言語の様な C のソースファイルに埋め込み、それを C コンパイルでコンパイルするというものである。perlcc を利用することで Perl インタプリタが無い状況でも可動するバイナリファイルを作成する事が可能である。しかし Perlcc は Perl スクリプトが複雑になるほど正確に C に移植を行う事が出来ず、現在では Perl のコアモジュールから外されている。Perlcc は Perl のバイトコードを C への変換のみ行う為、C で実装されている Perl 経由で実行した場合と処理速度はほぼ変わらない。また Perlcc で生成された C のソースコードは難解であり、これをデバッグするのが困難でもある。MoarVM で threaded code を実現出来た場合、その箇所のみ

CbC プログラムとして切り出す事が可能である為 perlcc と似たツールを作成することも可能である。C レベルでも Perlcc の様に内部構造を C の関数化すれば ThrededCode の様な物を構築できるが、CbC と比較して処理の単位が明確ではない為高速化は見込めない。CbC を用いた ThrededCode で Perlcc の様なツールを作成した場合、CodeGear の単位が正常に機能すれば CbC の CodeGear が ThrededCode をより効率化出来ると推測できる。

CbC の CodeGear は `goto` 文で遷移するため、次の CodeGear が一意に決定している場合 C コンパイラ側でインライン展開する事が可能である。CodeGear がインライン展開される限界については別途研究する必要があるが、CbC を利用した場合 CodeGear 単位でインライン展開が可能である。その為、ThrededCode を実装する場合に決定した次の CodeGear をインライン展開する事が可能である。従って ThrededCode を実現するにあたり新たな処理系を開発する必要がなく、既存の資源を利用して ThrededCode が実現出来る。これを繰り返す事で Perlcc などと比較してより高速化した ThrededCode が実現できる。

## 8. まとめ

本論文では CbC によって Perl6 の処理系である MoarVM インタプリタの一部改良とその手法を示した。CbC MoarVM ではオリジナルの MoarVM と比較して以下の様な利点が見られた。

- CodeGear 単位で命令処理を記述する事が可能となり、モジュール化が可能となった。
- ThrededCode を実装する際に効率的に実装ができる見込みが立った。
- CodeGear を導入した命令単位での最適化が可能となった。
- break point を命令の処理単位でかける事可能となった。

今後 CbC での開発をより深く行っていくにあたり、CbC コンパイラそのものの信頼性を向上させる必要がある。MoarVM の開発を行うにあたり新たに発見された複数のバグを修正し、より安定するコンパイラにする為に改良を行う。

現在 CbCMoarVM で直接バイトコードを入力した場合の nqp のテストは JVM, Javascript のテストを除く中で 80%パスする。また数値の計算と出力などの簡単な NQP の例題を作成し、オリジナルの NQP, MoarVM でバイトコード化したものを入力した際も正常に動作している。しかし NQP のセルフビルドは現在オブジェクトの生成に一部失敗している為成功していない。今後はさらに複雑な例題や NQP のセルフビルド、Perl6 の動作を行っていく。

MoarVM では GC からオブジェクトを守る為に MVMROOT というマクロを利用し、局所変数のポインタをスタックに登録する処理を行っている。GC の制御を効率的に行えば本来は必要ない処理であり、実行すると CodeGear の優位性が損なわれてしまう。従って MoarVM の GC の最適化を行う。

また高速化という面では、Perl の特徴である正規表現に着目し、正規表現の表現のみ高速で動く最適化の導入なども検討している。他に rakudo のコンパイラ系統から CbC のコードを直接生成させ、それを llvm でコンパイルすることによって LLVM の最適化フェーズを得て高速化することも可能であると推測できる。

Perl6 の開発は非常に活発に行われている為、CbCMoarVM の最新版の追従も課題となっている。現在は interp.c から Perl スクリプトを用いて自動で CbC の CodeGear を生成している。今後の開発領域の拡大と共により効率的に CbC コードへの自動変換も複数の C コードに対応する様に開発を行っていく。

## 参考文献

- [1] <http://luajit.org/>.
- [2] <https://github.com/perl6/nqp>.
- [3] <https://github.com/perl6/nqp/blob/master/docs/ops.markdown>.
- [4] <http://parrot.org/>.
- [5] <https://design.perl6.org/>.
- [6] <https://docs.perl6.org/>.
- [7] <http://hackage.haskell.org/package/Pugs>.
- [8] <http://edumentab.github.io/>

- [9] <http://edumentab.github.io/rakudo-and-nqp-internals-course/slides-day1.pdf>.
- [10] <https://github.com/perl6/roast>.
- [11] <https://www.complang.tuwien.ac.at/forth/threaded-code.html>.
- [12] titleThreaded Code (1973).
- [13] FabioRiccardiOptimizing direct threaded code by selective inlining (1998).
- [14] ShinjiKONOImplementing Continuation based language in LLVM and Clang (2015).
- [15] 河野真治 Gears OS のモジュール化と並列 API (2018).
- [16] 河野真治第 53 回プログラミング・シンポジウム (2012).
- [17] 河野真治琉球大学工学部情報工学科平成 27 年度学位論文 (修士) (2015).
- [18] 並木美太郎 Ruby 用仮想マシン YARV の実装と評価 (2006).