Szegedi Tudományegyetem

Informatikai Intézet

SZAKDOLGOZAT

Szűcs Dániel

2023

Szegedi Tudományegyetem
Informatikai Intézet

Es2panda fordítóprogram fejlesztések

Szakdolgozat

Készítette:
Szűcs Dániel
informatika szakos
hallgató

Belső Témavezető:
Dr. Kiss Ákos
egyetemi docens

Külső Témavezető:
Gál Péter
!TODO

Szeged
2023

Feladatkiírás

A témavezető által megfogalmazott feladatkiírás. Önálló oldalon szerepel.

Tartalmi összefoglaló

- ***A téma megnevezése:***

Es2panda fordítóprogram fejlesztése, az ETS nyelvi konformancia növelés érdekében.

- ***A megadott feladat megfogalmazása:***

A cél, hogy az Es2panda fordítóprogram funkcionalitását kibővítsem a különböző komponensei fejlesztése által. A Checker komponensnek képesnek kell lennie különböző, az objektum orientált programozási paradigma által leírt szemantikai elemzésekre, továbbá az ETS nyelv specifikus szkóp feloldási szabályainak kezelésére. A Compiler komponensnek képesnek kell lennie a feltételes vezérlési szerkezet és a különböző bináris operátorok bytecode-ra fordítására, valamint függvény szkópokban a paraméterek és lokális változók regiszter allokálására és spillelésére.

- ***A megoldási mód:***

Először is, a Compiler komponens egyszerűbb részein, a vezérlési szerkezeteken és az operátorokon, majd a Checker komponens OOP mintáinak szemantikai elemzésén dolgoztam, így átfogó ismeretet szereztem a komponensek felépítéséről. Ezek után a két komponenst párhuzamosan fejlesztettem, az új nyelvi elemek implementációs sajátosságainak megfelelően.

- ***Alkalmazott eszközök, módszerek:***

Az Es2panda fejlesztése Ubuntu 22.04 operációs rendszeren folyik, a C++ nyelv 17-es szabványában. A fejlesztéshez Visual Studio Code fejlesztői környezetet használok. A fordításhoz az LLVM toolchain 14-es verzióját használom, debuggoláshoz a cgdb eszközt választottam.

- ***Elért eredmények:***

Implementáltam egy robosztus rendszert a regiszter spillelésre, valamint a regiszterek statikus típusainak követésére, továbbá implementáltam a szkóp feloldást, amely felhasználóbarát fordítási hibaüzeneteket ad a fordítóprogram felhasználójának.

- ***Kulcsszavak:***

fordítóprogram, szemantikai analízis, regiszter spillelés

Tartalomjegyzék

1. Bevezetés

Szakdolgozatom témájául az Es2panda fordítóprogramon végzett fejlesztéseimet szeretném bemutatni. Mivel a fordítóprogramon végzett munkáim rendívül szerteágazóak, így sok különböző problémával találkoztam a fejlesztés során, melyekre a megoldás nem mindig volt triviális a már meglévő szoftver architektúra mellett. A fejlesztés alatt a következők voltak a céljaim:

- Robosztus könnyen érthető kód írása, amely követi a C++ nyelv bevált gyakorlatait. Ezek magukba foglalják az explicit memória kezelése minimalizálását, a const korrektséget, a makró használat minimalizálását, valamint a komponensek szoros összekapcsolódásának elkerülését.
- A futásidei teljesítmény növelése, akár a fordítási idő romlása árán. Ez leginkább a template metaprogramozási gyakorlatokkal értem el, melyek drasztikusan javítottak a kód hívó oldalának egyszerűsítésében.
- Az implementált függvények és típusok interfészeinek egyszerűsítése, hogy minél nehezebb legyen őket rosszul használni.
- A modern fordítóprogramok által alkalmazott technikák kutatása és az Es2pandába való integrálásuk. Ezek magukba foglalják a különböző optimalizációkat és statikus analíziseket.

A következő oldalakban, betekintést adok az Es2panda általam fejlesztett részeibe, azok tervezési folyamataiba, és példákat mutatok a teljesítmény orientált gondolkodásmódra, amely elengedhetetlen egy fordítóprogram esetében. Először is bemutatom az Es2pandát jelenlegi állapotában, majd egy áttekintést nyújtok az általános struktúrája felett. Bemutatom az Es2panda hat nagy komponensét és rajtuk végzett fejlesztéseimet, mindezt a komponensek függőségének sorrendjében, kezdve a belépéstől a kész bytecode-ig. A komponenseket és az Es2panda kódját az általa támogatott ETS nyelvet érintő részekből fogom bemutatni. Végezetül kitérek a generált bytecode-ra és az érdekesebb implementációs részletekre.

2. Mi az az Es2panda?

Ebben a fejezetben általánosan beszélek az Es2pandáról mint szoftverről.

2.1. Mi az a fordítóprogram frontend?

A fordítóprogram frontend egy olyan szoftver, amely egy előre meghatározott nyelvet vagy nyelveket egy köztes reprezentációra vagy más néven bytecode-ra fordítja, melyet később egy fordítóprogram backend tényleges gépiódra fordít, vagy egy interpreter futásidőben értelmez.

2.2. Mi az az Es2panda?

Az Es2panda egy fejlesztés alatt álló fordítóprogram frontend melyet C++ nyelven fejlesztünk. A program több különböző szkriptnyelvet támogat és párhuzamos fordítással rendelkezik.

2.3. Hogyan fejlesztjük az Es2pandát?

Az Es2panda fejlesztése Linux platformon történik, a projekt fordításához szükséges csomagokat egy bootstrap szkript tudja telepíteni Ubuntu 20.04-es és 22.04-es kiadásokra. A build rendszer a CMake 3.10-es verziója. A build konfigurálható gcc és clang fordítóprogrammal is, a támogatott verzió gcc esetében 11.1, clang esetében pedig 14.0. Az Es2panda 64 bites Unix alapó rendszerekre lefordítható, mind x86-64 mind arm64 architektúrára. A projekt a C++17-es nyelv szabványt használja.

A verzió kezelést biztosító szoftver a git. A forráskód megtalálható a Gitee weboldalon. A különböző toolchain komponensek, mint az Ark runtime, az Es2panda, és a runtime nyelvi pluginjai (Ecmascript, ETS) különböző repositorykban érhetők el.

3. Az Es2panda struktúrája

Az Es2panda mind logikailag, mind fizikailag szegmentálva van a struktúráját illetően. Ez a szegmentáció segít átlátni az absztrakciókat, kezelni a dependenciákat, elrejteni az implementációs részleteket és optimalizálni a fordítási időt.

3.1. A fizikai struktúra

Az Es2panda a *panda::es2panda* névtéren belül helyezkedi el, a kód példákban ezt a névteret nem fogom külön megnevezni. Az program a standard C/C++ header módszerrel van strukturálva. Az absztrakt szintaxis fa (AST) minden node-ja a saját header fájljában van, viszont a hat nagy logikai komponens javarészt monolitikusan egy-egy header fájlban található. Néhány függvény, amelyek különben implementálhatók lennének egy külön fordítási egységben (*.cpp*), mégis headerben van implementálva. Ennek az oka a fejlesztés elején a fordítási idő gyorsítása volt, valamint a hívások jobb optimalizálása fordítási egység határokon (*inlining)*. Ezeket a függvényeket a fejlesztés jelen szakaszában próbálom megszüntetni, ha egy módosításom érinti őket, ugyanis a támogatott fordítóprogramok már képesek a linkelési idejű optimalizációra (LTO), így a legtöbb belső szimbólum deduplikálásra kerül.

A projekt tartalmaz több olyan fájlt is, amelyeket a build rendszer fordítás előtt generál. Ezek a fájlok általában olyan kódot tartalmaznak, amelyek repetitívek, vagy könnyen generálhatók egyszerű markup nyelvekből. A projekt a YAML markup nyelvet használja, amit Ruby szkriptek dolgoznak fel, majd az Embedded Ruby (ERB) rendszer alakítja át megfelelő forrás fájllá, általában headerré. Néhány generált fájlra a logikai egységek részletezésénél külön ki fogok térni.

Az Es2panda két ELF állományból áll. Az egyik az *libes2panda-lib.so* programönyvtár, ami a tényleges fordítási logikát tartalmazza, a másik maga az *es2panda* nevű futtatható állomány, ami dinamikusan linekelt az előző programönyvtárhoz. Ez a szeparáció lehetővé teszi, hogy a fordítóprogram a futtatható állomány nélkül is integrálható lehessen más rendszerekbe. Azonban természetéből adódóan a szeparáció ront az LTO hatékonyságán, mivel a programkönyvtár szimbólumait nem lehet eltávolítani a dinamikus linkelés miatt.

3.2. A logikai struktúra

Ebben a fejezetben röviden bemutatom a logikai egységeket, a mélyebb működésükre és implementációs részleteikre a saját fejezeteikben fogok kitérni, kivéve az első egység esetében. Mivel ennek az egységnek a fejlesztésében nem foglaltam kulcsfontosságú szerepet, így nem szentelek neki külön fejezetet. Az Es2panda fordítóprogram hat nagy logikai elemre van strukturálva. Ezek név szerint a Lexical Analyser, a Parser, a Binder, a Semantical Checker, a Code Generator, és az Emitter. Van egy hetedik elem is, amin az előbb említett logikai elemek dolgoznak, ez az absztrakt szintaxis fa (AST).

3.2.1. A Lexical Analyser (Lexer)

A Lexical Analyser (Lexer) felelős a beolvasott forráskód tokenizációjáért. Az Es2panda indulásakor a Lexer az első komponens amelyik inicializálódik. A Lexer szolgáltatja a Parser számára a tokeneket és a sor, oszlop információt hibaüzenetek esetén. Ennek a komponensnek egy része, a kulcsszavak beolvasása egy YAML markup fájlból generálódik a build során.

A Lexerből a *Lexer::NextToken* függvénnyel lehet az aktuális tokent kikérni, ez előre is mozgatja a lexer fejét a következő tokenre. Lehetőség van az aktuális pozíción lévő tokent követő karakter megvizsgálására, ezt a *Lexer::Lookahead* függvény valósítja meg. Esetenként előfordul, hogy a parsolás során a Lexert vissza kell tekerni. Erre lehetőséget adnak a *Lexer::Save* és a *Lexer::Rewind* függvények.

A generált *keywords.h* nevű fordítási egység valósítja meg a tényleges tokenizációt. A generált fájl függvényei a *Keywords::Scan_** alakúak, ahol a ‘*’ a következő betűt jelöli. Például, ha az első függvény a *Keywords::Scan_i* volt, akkor a következő hívott függvény négy különböző lehet. Lehet a *Keywords::Scan_if* ami ha f-et whitespace követi felvesz egy ’if’ *Lexer::Token-t* ami kulcsszó. Hasonlóan lehet a *Keywords::Scan_im és Keywords::Scan_in* mely függvények az ’implements’ és ’interface’ kulcsszavakat próbálják szkennelni. Ha ezek közül egyik sem teljesül, akkor a negyedik esetben az előállított *Lexer::Token* azonosító (identifier) lesz.

A **Kód Részlet 1** mutatja hogyan lehet egy kulcsszót felvenni a leíró YAML fájlba.

```yaml
keywords:
- name: 'async'
  token: KEYW_ASYNC
  keyword: [as, ets]
  keyword_like: [js, ts]
```

Kód Részlet 1: ’async’ kulcsszó YAML rekord.

A *name* mező a kulcsszó konkrét formáját adja meg, amit a Lexer keresni fog a tokenizáció során. A *token* mező a *Lexer::Token-hez* tartozó *Lexer::TokenType* generált enum egyik mezőjének neve lesz. A *keyword* és *keyword_like* mezők megadják, hogy melyik nyelvben hogyan viselkedik ez a kulcsszó. Mivel az Es2panda fordítóprogram több nyelvet is támogat ezért a Lexer nagy része újra használható a nyelvek közös pontjai között. Tehát ahogy a **Kód Részlet 1** mutatja, az *async* kulcsszó az AssemblyScript és ETS nyelvek esetén hard kulcsszó, azaz minden kontextusban kulcsszó, míg Ecmascript és Typescript nyelvek esetén soft kulcsszó, azaz a megfelelő pozícióban kulcsszó más pozícióban pedig azonosító.

3.2.2. A Parser

A Parser logikai egység a Lexer által előállított tokenekből építi fel az absztrakt szintaxis fát. Mivel az implementációja egy LR(1) -es parser így ezt alulról felfelé teszi, először a levél node-okat állítja elő majd azok szülő node-jait, így haladva egészen a gyökér node-ig. A Parser hívja a Binder megfelelő függvényeit, hogy a létrehozott AST node-ok megfelelő szkópokba legyenek bekötve.

3.2.3. A Binder

A Binder komponens a Parser komponenssel együtt dolgozva építi fel a szkóp fát. A szkópok *binder::Variable* pointereket kötnek nevekhez egy *std::unordered_map-ben*. A *binder::Variable* igazából egy szimbólum mivel nem csak változóra mutathat, hanem típus névre is és a megfelelő AST node deklarációjára mutat. Minden egyes szkópban a feloldás úgy történik egyszerűbb esetben, hogy a szkóp fát alulról felfelé bejárjuk, mindig csak a szülőkön iterálva, és megkeressük név alapján a megfelelő *binder::Variable-t*.

3.2.4. A Semantical Checker

A Semantical Checker az Es2panda legkomplexebb logikai egysége. Feladata a kód szemantikai analízise. Ez Ecmascript esetében gyakorlatilag egy üres operáció, minden más támogatott nyelv esetén viszont több feladatot is el kell látnia. Az Es2panda által támogatott nyelvek az Ecmascript kivételével mind rendelkeznek valamilyen típus rendszerrel. A típusok által hordozott információból és a nyelvek által támogatott különböző programozási paradigmák szabályrendszeréből a fordítóprogramnak warning és error üzeneteket kell előállítania, ha valamilyen szemantikai szabály sértett. A Semantical Checker gyakorlatilag majdnem minden AST node-hoz rendel egy *checker::Type* példányt, amelyben leírja az adott típust és különböző típus relációkat tud felállítani. Például osztályok között inheritencia láncot tud felállítani, vagy primitív típusok közötti konverziókat tud a primitívek mérete alapján engedélyezni vagy letiltani.

3.2.5. A Code Generator

A Code Generator az absztrakt szintaxis fából, és a Semantical Checker által előállított típus információból generál bytecode-ot. Ez a bytecode statikusan típusos, kivéve Ecmascript esetén, ahol az Ark runtime egy plugin segítségével támogatja a dinamikus bytecode-ot. A Code Generator párhuzamosan fut a fordítási egységekben lévő függvényekre. Ezt a működést mutexek nélkül tudja végezni (kivéve a szál scheduling esetében). A bytecode gyakorlatilag egy regiszter gép. Egy regiszter az akkumulátor szerepét tölti be. A Code Generator felelős a regiszterek megfelelő kiválasztásáért a regiszter spillelésért. A Code Generator egy vectorban tárolja el a kigenerált utasításokat.

3.2.6. Az Emitter

Az Emitter komponens fordítja át az Es2panda belső reprezentációját az Ark runtime reprezentációjára. Az Emitter ezenfelül annotációkkal látja el a bytecode-ot. Ezeknek az információknak egy részét az Ark runtime más részét a debugger használja fel.

3.2.7. Az Absztrakt szintaxis fa (AST)

Az AST az Es2panda legfontosabb adat szerkezete. Megvalósítását tekintve egy osztály hierarchiából áll, melynek gyökere az *ir::AstNode* pure virtual base osztály. Ez az osztály három fontos adattagot deklarál. Az első az *ir::AstNodeType* enum, aminek segítségével egy makrón keresztül *Is##NodeType* és *As##NodeType* típus ellenőrző, és típus konverziós függvények kerülnek kigenerálásra. A második szintén egy enum típusú adattag, az *ir::ModifierFlags*. Ahogy a neve is mutatja ezek a flagek bizonyos modifier kulcsszavak meglétét tárolják egy AST node-on. Ilyenek például a Java-ból ismert és az ETS által is használt láthatóságot szabályozó kulcsszavak a *public, protected, private* vagy más egyéb módosítók például *abstract, async, const, constructor, static*. A harmadik legfontosabb adattag a szülő node-ra mutató AST node pointer.

Az *ir::AstNode* négy fontos pure virtual függvényt deklarál. Ezek az *Iterate, Dump, Check* és *Compile* függvények. A *Check* és *Compile* függvények külön overloadokkal rendelkeznek Javascript és ETS esetekre.

- Az *ir::AstNode::Iterate* virtuális függvény paraméterében egy callback függvényt vár, amelynek szignatúrája *void(ir::AstNode *)*. Ez a virtuális függvény egy node minden gyerekére meghívja a paraméterben kapott callback függvényt.
- Az *ir::AstNode::Dump* kírja az adott node JSON reprezentációját. A függvényt az előzővel együtt használja a program a teljes AST kiírására. Ez az osztály képes több különböző primitív, vagy tömb szerű adatszerkezetet JSON formátumba alakítani.
- Az *ir::AstNode::Check* függvény visszaadja egy adott node típusát. A típust a Checker komponens állítja elő. Általában a függvény mellékhatása, hogy a node típusát is beállítja, ha a node egy *ir::TypedAstNode*.
- Az *ir::AstNode::Compile* függvényt a Code Generator hívja. Ebben a függvényben az AST node elvégzi a lefordításához szükséges lépéseket, majd az eredményt az akkumulátorba teszi.

Az *ir::AstNode-nak* három fontosabb leszármazott osztálya van. Az első az *ir::Statement* ami az utasítások ősosztálya. A második az *ir::TypedAstNode* amely kibővíti a funkcionalitást típus információval. A másodikból örököl a harmadik *ir::Expression*, ami a kifejezések ősosztálya.

4. A Parser

A Parser komponens állítja elő az AST-t. Az Es2panda által használt parser egy LR(1) parser. Ez azt jelenti, hogy az AST-t alulról felfelé építi fel. A Parser komponens *parser::AllocNode* függvénye az *ir::AstNode::Iterate* függvény mechanizmusát használja az új node-ok létrehozására. Miután lefut az allokálni kívánt node konstruktora, a *parser::AllocNode* meghívja az *ir::AstNode::Iterate* függvényt egy olyan callbackkel, amely az allokált node összes gyerekének beállítja a szülőjét az új node-ra. Ez gyakorlatilag a Parser általános mechanizmusa.

4.1. A Parser belépési pontja

A Parser belépési pontja a *parser::Parser::ParseProgram* függvény. Ez a függvény ETS esetében először is felépíti az ETSGLOBAL nevű osztályt. Erre azért van szükség mert a bytecode nem támogatja a globális változókat és függvényeket, viszont az ETS nyelv igen, erre nyújt megoldást az ETSGLOBAL. Ez az osztály gyakorlatilag egy shadow, amelyre minden globális változó és függvény fel lesz véve, mint statikus property.

Következő lépésben a default import-ok kerülnek parsolásra. Erre azért van szükség, mert az ETS nyelv előírja, hogy minden fordítási egység impliciten importálja az *std.core* csomagot. Ebben a csomagban érhető el a Java-hoz hasonlóan az *std.core.Object* osztály, ami az osztály hierarchia gyökere. Ezek mellett még sok más funkcionalitást is biztosít.

Az utolsó lépésben a *parser::ETSParser::ParseETSGlobalScript* függvény hívódik. Ez a függvény először összegyűjti az egymodulba tartozó fájlokat, majd elindítja a parsolás lényegi részét a *parser::Parser::ParseTopLevelDeclaration* függvényt, amely addig olvassa a Lexer által előállított tokeneket ameddig el nem jut a *lexer::TokenType::EOS-ig*. Ezután allokálódik a AST gyökér node-ja az *ir::ETSScript*.

4.2. A parser komponensen végzett fejlesztéseim

A Parser komponensen végzett fejlesztéseim főként a default és named argumentumok parsolása. Ezek az ETS nyelvi elemek a Python-hoz hasonlóan lehetővé teszik, hogy egy függvényt opcionális argumentumokkal ruházzunk fel és az is lehetséges, hogy az

argumentumokat a függvény deklaráció argumentum listájától eltérő sorrendben adjuk át, expliciten megnevezve azokat.

4.2.1. A default argumentumok

A default argumentumokkal rendelkező függvények esetében nem csak egy, hanem kettő függvényt is elő kell állítani. Az eredeti függvény mellett a második egy wrapper függvény, aminek feladata, hogy runtime megvizsgálja, hogy ténylegesen mely argumentumok lettek átadva, és a hiányzó argumentumok default incializáló expressionjét meghívja.

Az első feladat létrehozni egy új AST node-ot. Ez az AST node az ir::*ETSParameterExpression* tartalmazza az információt minden függvény, minden argumentumáról. Gyakorlatilag minden *ir::ETSParameterExpression* két részből áll, amit az ETS nyelv nyelvtana diktál. Az egyik egy *ir::Identifier*, a másik pedig egy *ir::Expression*. Ezek közül a második az argumentum default initializere, ami opcionális. A **Kód Részlet 2** mutatja az AST node struktúráját.

```
class ETSParameterExpression : public Expression {
public:
    explicit ETSParameterExpression(Identifier *left,
Expression *right);

private:
    Identifier *name_;
    Expression *initializer_;
};
```

Kód Részlet 2: Az *ir::ETSParameterExpression* struktúrája

A getter függvényeket a kód részlet tömörségének érdekében kihagytam. Setter függvényekkel a node nem rendelkezik, mivel sem egy argumentum neve, sem default initializere nem változhat.

A következő feladat az új node parsolása. A default argumentum parsolása a meglévő parser infrastruktúrával egyszerű. Először is egy nevet, azaz *ir::Identifier-t* kell parsolni majd egy típus annotációt. Ha ezen a ponton hiányzik bármelyik is akkor szintaxis hibát dobok. Ezek után opcionálisan parsolni kell egy *ir::Expression-t*, ha az argumentumnak van default initializere. Ezt a lépés sorozatot addig kell ismételni ameddig van még argumentum a függvény deklaráció szignatúrájában.

4.2.2. A proxy függvény

Miután a teljes függvényt beparsoltam, elő kell állítani egy proxy függvényt, amely becsomagolja azokat a függvény hívásokat, amelyek használják a default argumentumok értékét. Egy default argumentum initializerét természetesen csak akkor kell kiértékelni, ha azt az argumentumot nem adják át a függvénynek. Ahhoz, hogy ez eldönthető legyen a függvény argumentumai közé be kellen szúrni még egy flaget amit a compiler állít be, hogy ténylegesen lett-e passzolva argumentum, ezt a flaget megvizsgálni és annak alapján kiértékelni a default initializert. Ez a logika rontana a függvény futás idején abban az esetben, amikor nincs használva default initializer. Ezért van szükség a proxy függvényre, amely pontosan ezt a logikát implementálja.

4.2.2.1. A proxy függvény struktúrája

A proxy függvény argumentum listája megegyezik a sima függvényével kivéve, hogy tartalmazza a fentebb említett flaget. A flag egy 32 bites változó, tehát 32 default argumentumot tud kódolni. Ha több default argumentumra van szükség, akkor egy újabb 32 bites flag változó kerül az argumentum listába. Miután a proxy függvény letesztelte a default argumentumokat és beállította az értékeket, áthív a tényleges overhead nélküli függvénybe. A proxy függvény minden szintetikusan előállított *ir::Identifier-e* mangled, tehát olyan név, amelyet a felhasználó sohasem tud leírni. Ez gyakorlatilag azt jelenti, hogy minden ilyen névbe egy ’#’ karakter van belefűzve. A compiler csak olyan hívásokat cserél a proxy hívására, ahol szükséges legalább egy default argumentum használata.

4.2.2.2. A proxy függvény parsolása

A proxy függvény parsolása természetesen nem tud úgy történni, mint egy sima függvény parsolása, mivel nincs jelen a forráskódban. Maga a proxy függvény létrehozásához nincs szükség semmilyen szkóp, vagy típus információra, ezért az egész szintetikus függvény létrehozható parse időben. Mivel kézzel felépíteni egy teljes függvény AST-jét komplex fenntarthatatlan kódhoz vezetne, ezért helyette a *parser::InnerSourceParser-t* használom. Ez az osztály gyakorlatilag kicseréli a Lexer komponenst az aktuális Parser alatt, és egy előre összeállított stringre cseréli a

forráskódot. A parsolás tehát úgy történik, hogy a proxy függvény ETS kódját előállítom egy *std::string-be* majd a Parser mintha csak egy sima függvényt látna, beparsolja és a Binder megfelelő függvényeivel létrehozza a szkópokat és beköti a változókat. A **Kód Részlet 3** mutatja a *parser::InnerSourceParser* létrehozását.

```
const InnerSourceParser isp(this);
GetContext().Status() |= ParserStatus::ALLOW_HASH_MARK;
const auto lexer = InitLexer({"<default_methods>.ets",
proxy_method});

auto *const ident = AllocNodeNoSetParent
<ir::Identifier>(util::StringView(proxy_method_name), Allocator());
ParseClassMethodDefinition(ident, method->Modifiers());
GetContext().Status() &= ~ParserStatus::ALLOW_HASH_MARK;
```

Kód Részlet 3: A *parser::InnerSourceParser* létrehozása.

A *parser::InnerSourceParser* paraméterében elkapja az aktuális *parser::ETSParser this* pointerét, ezzel a konstruktorában elmenti a Lexer állapotát. Következő lépésben a Parser státuszára felkerül a *ParserStatus::ALLOW_HASH_MARK* flag. Erre azért van szükség mert a parser alapvetően nem engedi a '#' karaktert használni string literálokon kívül. A *parser::InitLexer* függvény létrehoz egy új lexer-t. A *<default_methods>.ets* egy placeholder név, ha a parsolás során hiba keletkezik, akkor a Parser erre a dummy fájlra fog hivatkozni. A *proxy_method* a kézzel összeállított ETS kód. Ezután a *parser::Parser::AllocNodeNoSetParent* függvény allokál egy *ir::Identifier-t* a függvénynek majd ezzel az azonosítóval elindul a parsolás a *parser::Parser::ParseClassMethodDefinition* függvényben. Ez a függvény gyakorlatilag nem is látja, hogy éppen egy szintetikus függvényt parsol. A függvény végén leveszem a flaget a Parser státuszáról, majd a *parser::InnerSourceParser* destruktora visszaállítja az elmentett Lexert.

4.2.2.3. A megvalósítás összehasonlítása

A *parser::InnerSourceParser* használatának legfőbb előnye, hogy nem kellett a kódban ugyan azt a logikát kétszer leírni. Ha kézzel lenne felépítve az *ir::MethodDefinition* AST node, akkor ugyan azt a kódot kellene megismételni, mint egy hagyományos függvény parsolása esetén, kivéve a Lexer mozgatásokat. Mivel egy függvény potenciálisan több száz AST node-ból is állhat, ez a lehetőség komplex, hosszú, fenntarthatatlan és bugprone kódhoz vezetne.

5. A Binder

A Binder komponens építi fel a szkóp fát, és kezeli a szkópok közötti váltásokat. A Binder nyilván tartja az éppen aktuális szkópot és a globális szkópot egy-egy *binder::Scope* pointer adattag segítségével. A Binder komponens építi fel az AST node-ok teljesen kvalifikált neveit is mivel ehhez a minden névvel rendelkező szkópra szükség van.

5.1. A szkóp fa

A szkóp fa ősosztálya a *binder::Scope*. Hasonlóan az AST ősosztályához a *binder::Scope* is tartalmaz egy szülőre mutató pointert. Ezen kívül a szkóp tratalmaz egy *ir::AstNode* típusú adattagot is, ez általában egy *ir::BlockStatement* ami a szkóp létrejöttét eredményezte. A *binder::Scope* legfontosabb adattagja a *binder::VariableMap* ami igazából csak egy alias a egy *std::unordered_map-re*. Ez a map tartja számon, hogy egy szkópban milyen névhez, milyen változó, vagy típusnév tartozik. Továbbá, a map szemantikájából adódóan megoldja a változó nevek elrejtését is. Alapvetően a mapbe mindig az *std::unordered_map::insert_or_assign* C++17-es függvényt használva szúrunk be, ami kulcs ütközés esetén felülírja az értéket.

A *binder::Scope* feletti egyik legfontosabb absztrakció a *binder::LexicalScope*. Ez az absztrakció teszi lehetségessé Javascript esetén az *eval* ellenőrzését, azonban ETS szempontból nem sok jelentősége van, mivel ETS-ben nincs *eval*. Ennek ellenére a *binder::LexicalScope::Enter* statikus függvényt kell egy szkópba való belépéshez hívni. A belépés azt foglalja magába, hogy a *binder::LexicalScope* kicseréli a Binder szkóp pointerét a *binder::LexicalScope::Enter* által kapott pointerre. Ez azért fontos mert a Parser amikor egy *ir::Identifier-t* allokál, akkor nem tudja, hogy milyen szkópban áll, ezért mindig a Binder által mutatott aktuális szkóp *binder::VariableMap-jéhez* fűzi hozzá.

5.2. A Binder komponensen végzett fejlesztéseim

A Binder komponensen végzett fejlesztéseim, a szkóp keresés eredmény típusának a *binder::ScopeFindResult-nak* az újra implementálása és az azonosító feloldásnak az újra dolgozása volt.

5.2.1. A szkóp keresés eredmény típus újra implementálása

A szkóp keresés eredmény típusa a *binder::ScopeFindResult* típus egy egyszerű típus, amit a *binder::Scope-on* található kereső függvények adnak vissza. Ez a típus három fontos adattaggal rendelkezik:

- A feloldott név, amit a kereső függvények kaptak.
- Egy *binder::Scope* pointer, ez a szkóp az, amiben a kereső függvény megtalálta a keresett nevet.
- Egy *binder::Variable* pointer, ami a név alapján feloldott szimbólum.

A három változó közül a *binder::Scope* pointer problémásnak bizonyult több kereső függvény esetében is. Mivel a függvényekben előfordulhat, hogy a visszaadott *binder::ScopeFindResult binder::Scope* pointere az maga a *this* pointer, és ezt a pointert a *binder::ScopeFindResult* non-const pointerként tárolta ezért a kereső függvényeket nem lehetett konstanssá tenni. Ez a tény az egyszerű kereső függvények esetében hátrányos, mert a kiadott pointeren keresztül nem változtak a szkópok, viszont ezt a tényt a függvény nem kommunikálta. A típus rendelkezik továbbá még kettő *uint32_t* típusú adattaggal, amik csak Javascript esetben vannak használva.

5.2.1.1. A megvalósítás

A *binder::ScopeFindResult* típust először is átneveztem *binder::ScopeFindResultT-re* és template-sé tettem. A template-nek egyetlen típus paramétere van a *ScopeT*. Ezt a típus paramétert szerettem volna megszorítani C++20 *requires* klózokkal, viszont a projekt C++17-ben van írva. A megszorításokat így kénytelen voltam a C++17 által nyújtott *std::enable_if_t* típus template paraméter használatával megoldani. Az *std::enable_if_t* egy olyan template típus, ami kihasználja a SFINAE-t. Az *std::enable_if_t* feltétele, hogy a kapott *ScopeT* típus, pointer és *binder::Scope* vagy annak leszármazott típusa.

Visszatekintve, egy jobb implementáció lett volna, ha a megszorítások között a pointer típus nem szerepel, és csak az inheritencia vizsgálat szerepelne. Ez az átdolgozott *binder::Scope::Find* függvények implementációját is tisztábbá tette volna.

Végezetül készítettem két *using* deklarációt *binder::ScopeFindResult* és *binder::ConstScopeFindResult* néven, hogy a visszaadott típusok nevei kifejezőbbek legyenek.

5.2.1.2. Az általános kereső függvény megvalósítása az új típusokkal

Az általános kereső függvény működése rendkívül egyszerű. Gyakorlatilag a *binder::Scope* szülein iterál felfelé, ameddig meg nem találja a keresett deklarációt a szkópban. Ha nem találja meg akkor a *binder::ScopeFindResultT binder::Variable* pointere *nullptr* lesz. Ennek a függvénynek csinálnom kellett konstans én nem konstans overloadot is, mert esetenként a visszaadott *binder::ScopeFindResultT binder::Scope* pointerén keresztül, mutáció történik. Ennek ellenére nem szerettem volna megismételni az implementációt kétszer úgy, hogy ez a két overload között a különbség csak és kizárólag a visszatérési értékük.

Ennek elkerülésére létrehoztam a *binder::Scope::FindImpl* függvényt. Ez a függvény egy template függvény, amelynek két típus paramétere van. Az első a *ResultT* ezt az előbbiekhez hasonlóan korlátoztam két típusra az *std::enable_if_t* segítségével. Ez a két típus a fentebb említett *using* deklaráció által definiált *binder::ScopeFindResultT* template példányok. A második típus paraméter a *ScopeT* amire ugyan azok a megszorítások vonatkoznak, mint a *binder::ScopeFindResultT* típusparaméterére. A függvény a *binder::Scope* pointert forwarding referenciával veszi át, tehát a típus megtartja minden esetben a *const* és *volatile* dekorátorait, így lehetséges, hogy a *binder::Scope::Find* függvény konstans és nem konstans overloadja is ugyan ezt a függvényt hívják.

A **Kód Részlet 4** mutatja a *binder::Scope::FindImpl* függvény SFINAE template fejlécét.

```
template <typename ResultT, typename ScopeT,
   std::enable_if_t<std::is_same_v<ResultT,
ConstScopeFindResult> || std::is_same_v<ResultT,
ScopeFindResult>, bool> = true,
   std::enable_if_t<std::is_pointer_v<ScopeT>
&& std::is_base_of_v<Scope, std::remove_pointer_t<ScopeT>>,
bool> = true>
static ResultT FindImpl(ScopeT &&scope, const util::StringView
&name, const ResolveBindingOptions options);
```

Kód Részlet 4: A *binder::Scope::FindImpl* függvény fejléce

5.2.2. Az azonosító feloldás módosítása ETS esetben

Az azonosítók feloldása az ETS nyelvben a megszokottól egy kicsit másképpen működik. A szabályok rá a következők:

- Ha az azonosítót a *this* kulcsszó kvalifikálja, akkor a keresett azonosítót az aktuális osztályban és az ősosztályaiban, interfészeiben kell keresni, csak a nem statikus adattagok között.
- Ha az azonosítót a *super* kulcsszó kvalifikálja, akkor a keresett azonosítót az ősosztályokban és az ősinterfészekben kell keresni csak a nem statikus adattagok között.
- Ha az azonosítót az osztály neve kvalifikálja, akkor a keresett azonosítót az aktuális osztályban és az ősosztályaiban, interfészeiben kell keresni, csak a statikus adattagok között.
- Minden más esetben (egy eset maradt, a kvalifikálatlan azonosító), a keresett azonosítót először a lokális szkópban kell keresni kivéve, ha a lokális szkóp osztály szkóp, majd a globális szkópban.

Ezek a szabályok gyakorlatilag azt jelentik, hogy egy azonosítót akkor és csak akkor keresünk egy osztályban, hogyha valamilyen kvalifikálóval rendelkezik. Ez a hagyományos azonosító feloldástól merőben eltér. A hagyományos feloldás esetében elég csak a szkóp fán végig iterálni ameddig meg nem találjuk a keresett azonosítót.

Könnyű belátni ahhoz, hogy megtaláljuk, amit keresünk nem elég önmagában a szkóp fa. Szükség van a Semantical Checker által előállított típus információkra is mivel nem csak szülő szkópokban kell keresni az azonosítót, hanem egy potenciális inheritencia láncot is be kell járni, minden kvalifikált feloldás esetében. A Semantical Checkert érintő részeket az azt érintő fejezetben fogom részletesen taglalni.

A fenti okokból adódik, hogy ezt a feladatot nem lehet egészében a Binder komponensen belül implementálni, így ebbe a komponensbe csak kettő egyszerűbb új függvényt valósítottam meg. A továbbiakban ezek implementációit mutatom be.

5.2.2.1. A megvalósítás

A megvalósításban két új függvényt implementáltam. Az első a *binder::Scope::FindInGlobal* a másik pedig a *binder::Scope::FindInFunctionScope* függvény. Mind a kettő függvény konstans tagfüggvénye a *binder::Scope-nak.*

5.2.2.2. A globális szkópban kereső függvény

A globális kereső függvény, a globális szkópban keres. Ez azonban nem triviális keresés, mert ETS esetében kettő globális szkóp található. Az első globális szkóp a tényleges globális szkóp, ebbe a szkópba kerülnek a más modulokból importált nevek. Ilyen név például a standard függvénykönyvtárból automatikusan importált *Object*. A második az ETSGLOBAL nevű shadow osztály, amely az összes globális függvényt, típust és változót tartalmazza, ebből a fordítási egységből.

A függvény úgy működik, hogy egy while ciklusban feliterál addig a szkópig amelyiknek a szülő szkópja a tényleges globális szkóp, így az ETSGLOBAL szkópban fog állni először. Itt megpróbálja a feloldást végrehajtani, és ha sikerül akkor visszatér a feloldott értékkel. Ha nem sikerül akkor megpróbálja a feloldást végrehajtani a tényleges globális szkópban. Ha ez sem sikerül akkor a visszaadott *binder::ConstScopeFindResult binder::Variable* pointere *nullptr* lesz.

5.2.2.3. A függvény szkópban kereső függvény

A függvény szkópban kereső függvény implementációja valamivel egyszerűbb, mint a globális szkópokban kereső függvényé. Itt azonban azt kell figyelembe venni, hogy egy függvényben sok féle szkóp lehet például *binder::LoopScope*. Ezek a szkópok közül csak az osztály és a globális szkópokat kell kiszűrni.

Implementációját tekintve a függvény az aktuális szkóptól kezdeve iterál felfelé a szkópokon ameddig a következő szülő szkóp nem osztály vagy globális szkóp. Minden szinten történik egy kísérlet a feloldásra. A feloldás eredménye hasonlóan alakul a globális feloldó függvény eredményéhez. Ebben az esetben is egy *binder::ConstScopeFindResult* lesz az eredmény, aminek *a binder::Variable* pointere *nullptr* ha nem létezik a függvény szkópjában a keresett azonosító.

6. A Semantical Checker

A Semantical Checker komponens végzi a legnehezebb feladatot a fordítóprogramban. Feladata a statikus típus információk előállítása, statikus és szemantika analízisek elvégzése. Ilyen analízisek például az Objektum Orientált programozási paradigmából adódó öröklődési viszonyok ellenőrzése, virtuális függvény hívások feloldása, vagy lambda függvények capture változóinak ellenőrzése, és még sok egyéb fordítási idejű analízis.

6.1. A típus ősosztály

Maga a típus ősosztály a *checker::Type* nem rendelkezik sok adattaggal. Két adattagját érdemes említeni, az egyik egy *checker::TypeFlag* típusú tag, ami nevéből adódóan különböző flageket tárol egy típuson. Ilyen flagek például, hogy az adott típus primitív vagy referencia-e, esetleg szintetikus a Semantical Checker által előállított. A másik fontos adattagja egy *binder::Variable* pointer típusú adattag. Ezzel az adattagal csak olyan típusok rendelkeznek, amelyekhez valamilyen felhasználó által létrehozott típusnév is tartozik. Ilyenek például az osztályok, vagy a típus aliasok.

A típus ősosztály deklarál sok pure virtuális függvényt. Néhány egyszerűbb virtuális függvényei:

- *checker::Type::ToString* a típus string reprezentációja.
- *checker::Type::ToAssemblerType* egy string típust ad vissza, ami a felhasználó által deklarált típus assembler neve lesz a bytecode-ban.
- *checker::Type::ToAssemblerTypeWithRank* csak a tömb típusok használják, a tömb típusnevén túl reprezentálja a méretét is, különben úgy működik, mint az előző függvény

A típus ősosztály komplexebb virtuális függvényei a típusok relációival foglalkoznak, például:

- *checker::Type::Identical* eredménye igaz, ha a két típus megegyezik.
- *checker::Type::AssignmentTarget* megvizsgálja, hogy ha ez a típus egy értékadás bal oldalán áll, akkor a jobb oldalon álló kifejezés típusa kompatibilis-e vele.

- *checker::Type::Cast* megpróbálja a paraméterül kapott típust az aktuálisra castolni
- *checker::Type::IsSubType* megnézi, hogy a paraméterül kapott típus leszármazott típusa-e az aktuális típusnak

6.1.1. Az ETSObjectType típus

A *checker::ETSObjectType* típusról érdemes külön beszélni mert ez a legfontosabb előállított típus. Az ETS nyelv legtöbb típusa referencia típus a Javahoz hasonlóan, így minden típus, ami nem primitív, vagy enum ebből a típusból öröklődik.

A *checker::ETSObjectType* sok különböző adattaggal rendelkezik. *std::vector-ban* tárolja a következő tulajdnoságait:

- A típus argumentumait, ha generikus típusról beszélünk.
- A konstruktor szignatúráit.
- Az általa implementált interfészeket

Ezeken felül rendelkezik egy saját *checker::ETSObjectFlags* típusú flag adattaggal, ami a *checker::TypeFlags* felett még sok más tulajdonságot is beállít. A legfontosabb adattagja a *checker::ETSObjectType::PropertyHolder* nevű típus, ami a háttérben egy hat elemű tömb. A tömb minden eleme egy *std::unordered_map*. Ezek a mapek tárolják az összes propertyt egy osztályon. Ezek a propertyk lehetnek:

- Példány függvények.
- Statikus függvények.
- Példány adattagok.
- Statikus adattagok.
- Példány deklarációk, például inner belső osztály
- Statikus deklarációk., például statikus belső osztály

Ezen mapek nyilvántartása, és bennük az adott kontextusban a megfelelő property megtalálása a *checker::ETSObjectType* legfőbb feladata.

6.2. A check függvény

Az *ir::AstNode::Check* pure virtuális függvény az, amit a Semantical Checker használ. Ezt a függvényt hívja a gyökér node-tól kezdődően lefelé, mélységi bejárást alkalmazva. Előfordulhat azonban, hogy egy node típusa már egyszer létre lett már hozva, azonban újra ráfut a függvényre a vezérlés. Ebből az okból kifolyólag minden node függvénye először le ellenőrzi, hogy van-e már típusa, és ha van egyszerűen csak visszatér azzal.

6.3. A Semantical Checkeren végzett fejlesztéseim

A Semantical Checkeren számos fejlesztést végeztem, ezek közül a leglényegesebbeket szeretném bemutatni. Ez konkrétan három új működés implementálását jelentette, az első a beépített típusok típusinformációinak előállítása, a második az immutable adattagok egyszeri inicializálásának ellenőrzése, a harmadik pedig az **5.2.2-es** fejezetben említett azonosító feloldás.

6.3.1. A beépített típusok inicializálása

A beépített típusok típusinformációira mindig szükség van. Ilyen típusok például az *std.core.Object* ami az osztály hierarchia gyökere, vagy a *Console* objektum amely a standard kimenetre tud írni. Ezen típusok előállítása a megfelelő *ir::AstNode::Check* függvényeken keresztül történt, mint minden más node esetében is. Azonban ez a lazy módszer nem működött jól, sok helyen okozott problémát, hogy egy típus már inicializáltnak tekintette a *checker::GlobalTypesHolder-ben* lévő beépített típusokat.

6.3.1.1. A megvalósítás

A megvalósítás a *checker::InitializeBuiltins* nevű függvény implementálása. Ezt a függvényt hívja a *checker::StartChecker* függvény, ami a Semantical Checker belépési pontja. Az implementált függvény első lépése, hogy megvizsgálja, hogy a beépített típusok inicializálva vannak-e már. Ez egy biztonsági lépés mivel minden fordítási egységnek saját Semantical Checkere van, viszont a beépített típusokat elég egyszer megvizsgálni. A következő lépésben a globális szkópban megkeresem a beépített *std.core.Object* osztályt majd inicializálom. Azért kezdek ezzel a típussal, mivel minden

más típusnak be van kötve, mint ős, így garantált, hogy más típusok vizsgálata során ez a típus már előállt. Ezután végig iterálok a globális szkóp minden interfész és osztály típusán és egyesével beinicializálom őket a *checker::GlobalTypesHolder::InitializeBuiltin* függvényével. Ez a függvény megkeresi a beépített típusok között az adott típust, majd a kiszámított típusinformációt elmenti. Előfordulhat, hogy a fejlesztés során a standard függvénykönyvtárba olyan típus kerül be, ami még nem létezik a *checker::GlobalTypesHolder-ben*, ekkor, ha az Es2panda debug módban van buildelve akkor a loggeren keresztül egy figyelmeztetést adok ki, hogy az új típust fel kellene venni a beépített típusok közé.

6.3.2. Immutable adattagok egyszeri inicializálása

Egy osztály immutable adattagjait pontosan egyszer lehet inicializálni. Ezt az inicializációt két féle képpen lehet elvégezni, vagy az adattag deklarációjával egy időben (inline), vagy attól függően, hogy statikus vagy példány adattag az osztály static blokkjában, vagy konstruktorában.

6.3.2.1. A megvalósítás

A megvalósításom erre a problémára a következő. Először is végig iterálok az összes immutable addatagján az osztálynak a *checker::ETSChecker::CheckConstFields* függvényben. Második lépésben, ha az adott adattag statikus, akkor összegyűjtöm az osztályon lévő összes static blokkot, ha példány akkor pedig az összes konstruktort. Példány esetben megnézem, hogy a konstruktor első utasítása az egy másik konstruktorba delegáló hívás-e, mert ha igen akkor itt a változót inicializáltnak lehet tekinteni. Tovább menve végig iterálok a konstruktor vagy static blokk teljes utasítás listáján, és ha találok még egy elemet, ami inicializálná az adattagot, akkor hibát dobok.

A megvalósításnak van egy hibája, hogy nem veszi figyelembe a control flowt. Így előfordulhat a kódban olyan hely, ahol az inicializáló utasítás elérhetetlen ágon van. A fordított eset is előállhat, amikor hibásan dob hibát mert a második inicializáló utasítás van elérhetetlen ágon. Sajnos a projekt jelen állásában még nincs Control Flow Graph (CFG) építve, így ezeket a problémákat nem tudtam kijavítani.

6.3.3. Az azonosító feloldás módosítása ETS esetben

Az **5.2.2-es** fejezetben ismertettem az ETS nyelv azonosító feloldási szabályait, és bemutattam, hogy ehhez milyen fejlesztéseket kellett elvégezzek a Binder komponensben. Most a Semantical Checkert érintő módosításokról fogok beszélni.

6.3.3.1. A megvalósítás

A megvalósítás során két esetet kellett figyelembe vegyek.

Az első eset amikor egy azonosító önmagában áll, kvalifikáló kifejezés nélkül. Ebben az esetben először is megpróbálom feloldani a változót a függvény szkópjában. Ha ez nem sikerül akkor újra próbálom a feloldást a globális szkópban. Ez után a *checker::ValidateResolvedIdentifier* függvényt hívom. Ez a függvény végzi a feloldott változó validálását. Az első lépésben ellenőrzi a függvény, hogy sikerült-e a feloldás, és ha nem akkor megpróbálom feloldani az aktuális osztály szkópjában. Ezt azért teszem, hogy a felhasználónak jobb fordítási idejű hibaüzenetet tudjak adni. Például a felhasználó egy osztály adattagot szeretett volna elérni, de elfelejtette kirakni a *this* kulcsszót, akkor a hibaüzenet megmutatja neki, hogy a hivatkozott változót milyen kvalifikáló kifejezésen keresztül lehet elérni. A további vizsgálatok a feloldott azonosító által hivatkozott típusra, vagy változóra vonatkoznak, hogy az adott kontextusban helyes-e használni.

A második eset amikor egy kvalifikáló kifejezéssel rendelkező azonosítót kell megvizsgálni. Ebben az esetben először is össze kell állítani a jelenlegi kontextusból, hogy az osztályban adattagot, függvényt vagy deklarációt keresünk és az statikus-e vagy sem. Ennek a műveletnek az eredménye egy *checker::PropertySearchFlags* típusú flag lesz, amivel a *checker::ETSObjectType* property-jei között lehet keresni. Ezután megtörténik a feloldás. Az első esethez hasonlóan itt is validálva van a feloldott property, ezt a *checker::ValidateResolvedProperty* végzi. Ez a függvény megpróbálja a propertyt úgy feloldani, hogy ha eddig statikust keresett akkor most példányt és fordítva. Ez szintén a jobb hibaüzenet érdekében történik. A függvénynek ehhez meg kell fordítani a kereső flaget. A flagben hat különböző bit van, amik a keresett property típusokat kódolják. Ezt a hat bitet kell egymással kicserélni. A **Kód Részlet 5** mutatja a cserélő algoritmust.

```
unsigned int i, j; // positions of bit sequences to swap
unsigned int n;    // number of consecutive bits in each
sequence
unsigned int b;    // bits to swap reside in b
unsigned int r;    // bit-swapped result goes here

// XOR temporary
unsigned int x = ((b >> i) ^ (b >> j)) & ((1U << n) - 1);
```

Kód Részlet 5: A bit kicserélő algoritmus.

Mivel ez az algoritmus általános, viszont az én esetemben az *i* és *j* pozíciók, valamint az *n* konstans, így a művelet a következőkre redukálódik: **Kód Részlet 6**.

```
using Utype = std::underlying_type_t<PropertySearchFlags>;
const auto flags_num = static_cast<Utype>(flags);
const Utype x = (flags_num ^ (flags_num >> 3U)) & 7U;
const auto new_flags = PropertySearchFlags {flags_num ^ (x | (x
<< 3U))};
```

Kód Részlet 6: Az egyszerűsített bit cserélés

7. A Code Generator

A Code Generator feladata a típusokkal ellátott AST-ből, bytecode-ot generálni. Ehhez kettő építőelemre van szüksége. Az első a *compiler::IRNode*. Ez az osztály reprezentál a bytecode-ban egy utasítást. A másik fontos építőelem a *compiler::VReg*. Ez az osztály reprezentál egy allokált regisztert. A Code Generator az Es2panda egyetlen párhuzamosan futó komponense, minden egyes lefordított függvényhez létrejön egy példánya.

7.1. A Panda Bytecode

Mielőtt rátérhetnék a fent említett két elemre, beszélni kell az Es2panda által generált bytecode-ról. Ez a bytecode alapvetően egy regiszter alapú gép, viszont van egy kitüntetett regisztere az akkumulátor, amit a legtöbb utasítás impliciten használ. A Panda Bytecode típusos, a legtöbb utasítás több különböző típusra is elérhető. Néhány példa a teljesség igénye nélkül:

- *mov v1, v2* egy 32 bites értéket mozgat *v2-ből v1-be*
- *mov.64 v1, v2* egy 64 bites értéket mozgat *v2-ből v1-be*
- *mov.obj v1, v2* egy referenciát mozgat *v2-ből v1-be*

Az említett *mov* és változatai egy elég általános utasítás, mivel az operandusai típusát nem köti szigorúan, csak a méretüket. Az *fmovi* utasítás már szigorúbb, mivel ez egy 32 bites konstans lebegőpontos számot másol a cél regiszterbe.

A Panda Bytecode feltételezést tesz, hogy egy függvény argumentumait kevesebbszer használja, mint a lokális változóit, ezért az alacsonyabb regiszter indexeket a lokális változóknak osztja ki, míg a magasabbakat a paramétereknek. Ez a kiosztás lehetővé tesz egy optimalizációt a bytecode-on mégpedig, hogy a kevésbé használt utasításokat két bájton tudja kódolni, ez a rövid utasítás formátum. Nyolc bit kódolja az operációs kódot, a fennmaradó nyolc bit pedig vagy egy, vagy két regisztert kódol. Ez azt jelenti, hogy egy rövid utasításnak egy regiszteres esetben a maximum regiszter indexe kétszázötvenhat, két regiszteres esetben pedig tizenhat.

A normál formátumú utasítások legalább három bájton kódoltak, és általában két maximum kétszázötvenhatos indexű regisztert tudnak címezni, ez alól kivétel a *mov* utasítás, ami 2^{32} indexig képes címezni mindkét regiszterét.

7.2. Az IRNode

A *compiler::IRNode* osztály létesít absztrakciót a Panda Bytecode utasításai felett. Ez az osztály két pure virtuális függvénnyel ír le egy utasítást. Az első függvény a *compiler::IRNode::Registers* visszaadja az utasítás által használt regiszterek számát, és egy tömbben a regisztereit. A másik a *compiler::IRNode::GetFormats* függvény az utasítás formátumait adja vissza, azaz mekkora indexű regiszter operandusokat képes fogadni.

A *compiler::IRNode* tartalmaz továbbá egy *compiler::IRNode::Transform* függvényt, ami az Es2panda belső reprezentációjáról átalakítja az utasítást a runtime nyújtotta reprezentációra.

7.3. A VReg

A *compiler::VReg* egy vékony absztrakciós réteg a regiszter indexelés felett. Gyakorlatilag csak aritmetikai műveletek valósít meg, és specializálja a *std::hash* osztályt, hogy asszociatív adatszerkezetekben is tárolható legyen.

7.4. A Code Generátoron végzett fejlesztéseim

A Code Generatoron végzett fejlesztéseim a regiszter spillelés ETS nyelv esetén, valamint a regiszter allokálás implementálása volt.

7.4.1. A regiszter allokálás

A regiszter allokálás legtrükkösebb része a Panda Bytecode által megszorított allokálási stratégia betartása. A stratégia, ahogy a **7.1-es** fejezetben is említem a lokális változók regisztereit az alacsonyabb indexeken, míg az argumentumok regiszterei a magasabb indexeken foglalódnak, közvetlen a lokális változók után.

Az ETS nyelv viszont sok olyan konstrukcióval rendelkezik, amihez ideiglenes operációkat végezni. Ahhoz, hogy ezeket elvégezzük az operációk eredményeit regiszterekbe kell menteni. Ahhoz, hogy el tudjuk menteni őket ki kell számolni a következő regiszter indexet. Ehhez tudnunk kell, hogy mennyi lokális regisztert használunk. Nos, ez egy körkörös dependencia.

local n	local 1	local 0	arg 0	arg 1	arg n
65535 - n	65534	65535	65536	65537	65536 + n

Táblázat 1: A regiszter kiosztás

Ahogy a **Táblázat 1** mutatja a regiszter kiosztás a következőképpen történik a Code Generator futása során. A lokális változók *UINT16_MAX-tól* indulnak és lefelé nőnek, ez lehetőséget ad 65535 darab lokális változóra. Az argumentum regiszterek 65536-tól indulnak és tartanak *UINT32_MAX-ig*. A gyakorlatban sem a lokális, sem az argumentum regiszterek nem érik el az értéktartományaik maximumát.

Szóval a kód generálás során a folyamatosan allokált regiszterek lefelé nőnek, és minden spill-fill utasítás azonnal kigenerálásra kerül. Amikor a Code Generator terminál a regisztereket újra rendezi. A lokális regiszterek megtükröződnek az *uint16_t* értéktartományban, a paraméter regiszterek pedig redukálódnak *UINT16_MAX* mínusz az összes regiszter számával, így a felhasznált regiszterek teljesítik a Panda Bytecode által előírt allokálási stratégiát.

7.4.2. A regiszter spiller

A regiszter spilleléshez szükség van típus információra a regiszterekről ahhoz, hogy típusos bytecode-ot lehessen emittálni. A regiszterek típusát a *compiler::CodeGen* példány vezeti egy *std::unordered_map-ben*, aminek a kulcsa a *compiler::VReg*, értéke pedig a *compiler::CodeGen::VRegType.*

A *compiler::CodeGen::VRegType* reprezentálja a három állapotot, amiben egy regiszter típusa állhat. Egy regiszter típusa lehet *nullptr*, lehet egy konkrét típus, vagy lehet neki jövőbeli típusa. A jövőbeli típus azt jelenti, hogy a regiszter egy *out* regiszter, azaz a következő utasítás eredményként fog értéket kapni. A jövőbeli típus nem lehet *nullptr*, ezt az osztály interfész meg is követeli. A háttérben az osztály az *std::variant* típus biztonságos unió típussal van implementálva. Ezzel gyakorlatilag a *compiler::CodeGen::VRegType* is egy unió típus, ahol az unió elemei a három lehetséges állapot. A **Kód Részlet 7** mutatja a *compiler::CodeGen::VRegType* interfészét.

A típus információk birtokában már meg lehet kezdeni a regiszter spillelést. A *compiler::RegAllocator* minden egyes kiemittált utasítás előtt megnézi, hogy az utasítás regiszterei megfelelnek-e az utasítás formátumainak követelményeihez, ez a *compiler::RegAllocator::CheckRegIndicies* függvényben történik. Ha a regiszterek nem

felelnek meg az utasítás formátumainak, akkor a *compiler::RegSpiller* megkezdi a regiszter spillelést. Az ETS nyelv regiszter spillelését a *compiler::StaticRegSpiller* végzi. A spiller működése a következő, először is allokál egy új regisztert a nulla és tizenöt közötti 4 bites értéktartományban, ha ennek az új regiszternek már van típusa, azaz értéke is, akkor átmozgatja a következő üres regiszterbe. Ezen mozgatásokat feljegyzi a Spiller egy *std::vector-ba* mivel az új regiszter értékét majd vissza kell állítani. Ezután a használt regisztert átmozgatja az új regiszterbe, ha van típusa. Ha out regiszterről beszélünk, aminek jelenleg még nincs típusa, akkor a spiller feljegyez egy mozgatást az *std::vector-ba,* annak ellenére, hogy nem történt ilyen mozgatás. Erre azért van szükség mert a hívó kód az általa allokált regiszterben várja az értéket, az utasítás viszont csak a [0..15] intervallumon tud dolgozni, ezért az utasítás után rá kell venni a Spillert, hogy mozgassa ki az értéket az out regiszterből a hívó által allokáltba.

A *compiler::RegAllocator* kicseréli a *compiler::IRNode* regisztereit az újonnan allokáltakra. Ezután a Spiller visszafelé bejárja, az *std::vector-ját* és visszamozgatja a regisztereket eredti helyükre. Az **Ábra 1** mutatja a regiszter spillelés folyamatát.

8. Az Emitter

Az Emitter komponens gyakorlatilag egy adapter réteg az Es2panda és az Ark runtime Program reprezentációi között. Az Emitter komponens két alkomponensből áll. Az egyik a Record Emitter a másik pedig a Function Emitter.

A Record emitter felelősség a létrehozott osztályok emittálása. Az osztályokat a Rekord emitter a Binder komponensből gyűjti össze.

A Function emitter felelőssége a létrehozott függvények emittálása. A Function Emitter a Code Generator után fut egyből, tehát amint egy CodeGen példány lefordított egy függvényt, a Function Emitter átfordítja az Ark runtime reprezentációjára.

8.1. Az Emitter komponensen végzett fejlesztéseim

Az Emitter komponensen a generikus osztályok és függvények annotáció emittálását fejlesztettem. A generikus osztályok és függvények típus paraméterei rendelkezhetnek felső korlátokkal. Minden egyes kiemittált rekordhoz tartozik egy annotáció, ha generikus osztály volt, ami felsorolja a típus paramétereit és a felső korlátjaikat, ha van nekik. Ebből az információból később a debugger tud majd dolgozni.

Kidolgoztam továbbá egy adapter réteget, ami képes az Es2panda osztály reprezentációján használt flageket átfordítani az Ark runtime flagjeire.

A megvalósítás során az annotációkat generáló függvényeket felkészítettem arra is, hogyha a Java-hoz hasonló *@interface* annotációkat tudjon kezelni.

9. Összefoglaló

Összességében a fejlesztéseim az Es2pandában hasznosnak bizonyultak. Sikerült olyan rendszereket felállítsak, amelyek könnyen érthetőek és robusztusok. Célom volt a fejlesztések során az érintett kódok const korrektségének javítása, és a modern C++ legjobb gyakorlatok betartása, amely helyenként teljesítmény növekedéshez is vezetett, azonban a nagyobb jelentőség az előre észrevett bugok tekintetében van.

Természetesen mindig van lehetőség tovább fejlesztésre. A Code Generator komponens interfésze például túl gyenge megszorításokat tesz a Javascript és az ETS Code Generatoraira, ezért érdemes volna virtualizálni ezt az interfészt.

Sok helyen lenne érdemes jobban felbontani a Javascript és az ETS nyelvi elemeket kezelő kódok interfészeit a jobb kód fenntarthatóság érdekében. Ez kifejezetten igaz a Binder komponensre és függőségeire. A Binder komponens jelenleg majdnem ugyanúgy működik ETS esetén is mint Ecmascript esetén. Ez nyilvánvalóan nem tesz jót a kódfenntarthatóságnak. Vannak olyan szabályok is mint például az azonosító feloldás, melynek procedúrája teljesen máshogy működik a két nyelv között mégis egy azonos komponensen fut át. Ebben a komponensen rengeteg realizálatlan fejlesztési lehetőség van.

A Semantical Checker komponenst érdemes volna generációkra bontani a jelenlegi mélységi bejárás helyett. Ennek az lenne az előnye, hogy az AST node-okon elvégzett operációk sorrendje jobban definiált lehetne. A felbontás további előnye, hogy a Semantical Checker komponens túl monolitikus, a fenntarthatatlanság határán van, ezért átláthatóbb folyamatokhoz vezetne egy jobb nomenklatúra és esetleges Checker osztály hierarchia.

A Parser komponenst is lehetne tovább fejleszteni, bár messze ez a legstabilabb komponens a projektben, nagy fordítási egységek esetén ez a komponens sebessége határozza meg a teljesítmény maximumot. Ennek javítása érdekében lehetne a komponenst párhuzamosítani.

Irodalomjegyzék

[1] Sean Eron Anderson, Bit Twiddling Hacks, Swapping individual bits with XOR. Legutolsó látogatás: 2023. 05. 08.

[2] Ark Compiler Runtime Core. Legutolsó látogatás: 2023. 05. 08.

[3] Es2panda. Legutolsó látogatás: 2023. 05. 08.

Nyilatkozat

Alulírott Szűcs Dániel mérnökinformatikus BSc szakos hallgató, kijelentem, hogy a dolgozatomat a Szegedi Tudományegyetem, Informatikai Intézet Szoftverfejlesztés Tanszékén készítettem, mérnökinformatikus BSc diploma megszerzése érdekében.

Kijelentem, hogy a dolgozatot más szakon korábban nem védtem meg, saját munkám eredménye, és csak a hivatkozott forrásokat (szakirodalom, eszközök, stb.) használtam fel.

Tudomásul veszem, hogy szakdolgozatomat / diplomamunkámat a Szegedi Tudományegyetem Diplomamunka Repozitóriumában tárolja.

2023.

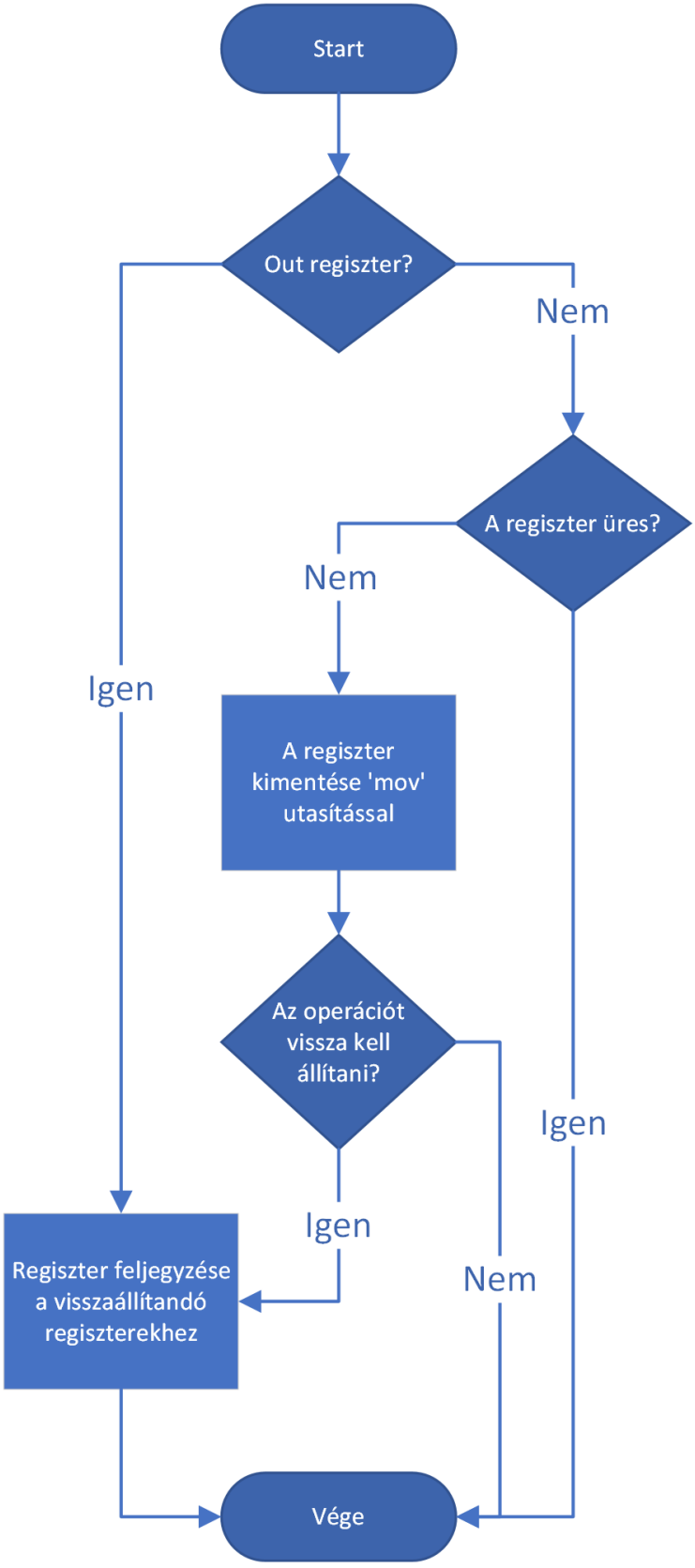
Szűcs Dániel

Mellékletek

Kód Részletek és Ábrák

```
class VRegType final {
    using Variant = std::variant<const checker::Type *, const
checker::Type *>;

public:
    [[nodiscard]] static constexpr VRegType FromType(const
checker::Type *type) noexcept;
    [[nodiscard]] static constexpr VRegType FromFutureType(const
checker::Type &future_type) noexcept;

    [[nodiscard]] constexpr std::optional<const checker::Type *>
GetType() const noexcept;
    [[nodiscard]] constexpr std::optional<const checker::Type *>
GetFutureType() const noexcept;

private:
    constexpr explicit VRegType(Variant &&types) noexcept;

    Variant types_ {};
};
```

Kód Részlet 7: A *compiler::CodeGen::VRegType* deklarációja.



Ábra 1: A regiszter spiller folyamatábrája