ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναφορά Εξαμηνιαίας Εργασίας

στο μάθημα 8ου εξαμήνου (ροής Λ)

Μεταγλωττιστές

Θέμα εργασίας: “*Ανάπτυξη ενός μεταγλωττιστή για τη γλώσσα προγραμματισμού Alan*”

Στοιχεία ομάδας

Ονοματεπώνυμο:	ΜΑΘΙΟΥΛΑΚΗ ΕΛΕΝΗ
Α.Μ.:	03114040
E-mail:	el.mathioulaki@gmail.com

Ονοματεπώνυμο:	ΝΙΑΡΧΟΣ ΣΩΤΗΡΙΟΣ
Α.Μ.:	03114076
E-mail:	sot.niarchos@gmail.com

Εισαγωγή

Αντικείμενο της εργασίας μας είναι η ανάπτυξη ενός μεταγλωττιστή για τη γλώσσα προγραμματισμού Alan, μία απλή προστακτική γλώσσα προγραμματισμού που περιγράφεται αναλυτικά στην εκφώνηση της εργασίας.

Η παρούσα αναφορά χωρίζεται σε αρκετά μέρη, ένα για κάθε τμήμα του μεταγλωττιστή. Συγκεκριμένα, η παρουσίαση θα γίνει ως εξής:

1. Εποπτεία της συνολικής αρχιτεκτονικής του μεταγλωττιστή
2. Λεκτικός αναλυτής (lexer)
3. Συντακτικός αναλυτής (parser) και κατασκευή αφαιρετικού συντακτικού δέντρου (AST)
4. Σημασιολογικός αναλυτής
5. Παραγωγή ενδιάμεσου κώδικα και βελτιστοποίηση
6. Παραγωγή τελικού κώδικα και βελτιστοποίηση
7. alanc – ο τελικός μεταγλωττιστής

Οι τεχνολογίες που χρησιμοποιήθηκαν σε κάθε τμήμα του μεταγλωττιστή είναι οι εξής:

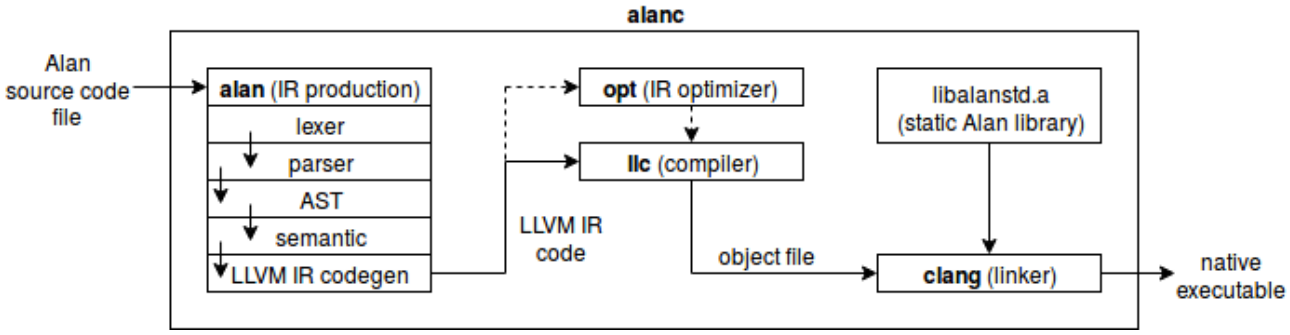
- Για τον λεκτικό αναλυτή χρησιμοποιήθηκε το εργαλείο flex
- Για τον συντακτικό αναλυτή χρησιμοποιήθηκε το εργαλείο bison
- Ο σημασιολογικός αναλυτής και η κατασκευή του AST γράφτηκαν σε C++
- Η παραγωγή ενδιάμεσου κώδικα έγινε σε C++ με χρήση των αντίστοιχων LLVM bindings (ο μεταγλωττιστής ελέγχθηκε με τις εκδόσεις 3.8 και 6.0 του LLVM)
- Η βελτιστοποίηση του ενδιάμεσου κώδικα, καθώς και η παραγωγή του τελικού κώδικα και η βελτιστοποίησή του έγιναν με χρήση κάποιων command line utilities που προσφέρονται από το LLVM (αναλυτικότερη αναφορά σε αυτά θα γίνει στη συνέχεια)
- Η standard βιβλιοθήκη της Alan γράφτηκε από την αρχή σε C, ακολουθώντας όσο πιο πιστά γινόταν αφενός την περιγραφή των συναρτήσεων στην εκφώνηση της εργασίας, αφετέρου τη συμπεριφορά των αντίστοιχων συναρτήσεων της C. Αυτό επιλέξαμε να το κάνουμε στα πλαίσια της προσπάθειας για μία πιο cross-platform προσέγγιση
- Το τελικό script που αλληλεπιδρά με τον χρήστη και συνενώνει όλα τα παραπάνω γράφτηκε σε Python 3.7. Η Python προτιμήθηκε ως “γλώσσα συνένωσης” (glue language) έναντι της αρχικής μας υλοποίησης σε bash με στόχο το τελικό script να είναι όσο το δυνατόν πιο εύχρηστο, ευανάγνωστο, επεκτάσιμο και cross-platform

Για το build του project αρκεί η εκτέλεση της εντολής make στο root directory (αυτό αποτελεί και το μόνο σημείο της εργασίας μας που, δυστυχώς, δεν συμμορφώνεται με την ιδέα ενός πλήρως cross-platform project).

1. Εποπτεία της συνολικής αρχιτεκτονικής του μεταγλωττιστή

Αρχικά, θα αναφερθούμε στην συνολική αρχιτεκτονική του μεταγλωττιστή, ώστε να είναι ξεκάθαρα η θέση και ο ρόλος του κάθε τμήματος το οποίο θα αναλύουμε στη συνέχεια της αναφοράς.

Μία αφαιρετική οπτική της αρχιτεκτονικής του μεταγλωττιστή που υλοποιήσαμε φαίνεται στο διάγραμμα που ακολουθεί. Πριν προχωρήσουμε στην επιμέρους ανάλυση των τμημάτων του μεταγλωττιστή, θα αναφερθούμε συνοπτικά στον τρόπο που αυτά αλληλεπιδρούν και ποιές είναι οι αρμοδιότητες του κάθε τμήματος, όπως αυτές προκύπτουν από τον σχεδιασμό μας.



Η καρδιά του μεταγλωττιστή μας είναι το πρόγραμμα **alan**. Αυτό είναι υπεύθυνο για την ανάγνωση του πηγαίου κώδικα Alan, την επεξεργασία του και, εν τέλει, την παραγωγή του αντίστοιχου ενδιάμεσου κώδικα, σε LLVM IR. Για να το επιτύχει αυτό, αποτελείται μεταξύ άλλων από τον λεκτικό, τον συντακτικό και τον σημασιολογικό αναλυτή. Ο λεκτικός αναλυτής αναγνωρίζει και προωθεί στο pipeline του alan τα lexemes, τα οποία στη συνέχεια ο συντακτικός αναλυτής χρησιμοποιεί για να κατασκευάσει το abstract syntax tree (AST) που αντιπροσωπεύει το αρχικό πρόγραμμα. Αυτό επιτυγχάνεται, προφανώς, στην περίπτωση που δεν εντοπιστούν ούτε λεκτικά ούτε συντακτικά λάθη. Στη συνέχεια, ο σημασιολογικός αναλυτής ελέγχει ολόκληρο το AST για σημασιολογικά λάθη. Εάν τέτοια δεν εντοπιστούν, ένα ακόμα τελευταίο πέρασμα του AST λαμβάνει χώρα, κατά το οποίο παράγεται ενδιάμεσος κώδικας στη γλώσσα του LLVM (LLVM IR), με χρήση των κατάλληλων C++ bindings.

Σε αυτό το σημείο, έχει χαθεί οτιδήποτε σχετίζεται με την Alan: έχουμε αυτή τη στιγμή στα χέρια μας το ισοδύναμο του αρχικού προγράμματος σε LLVM IR, το οποίο σημαίνει πως τώρα μπορούμε να το αντιμετωπίσουμε ως τέτοιο, απολύτως ανεξάρτητα από την αρχική (πηγαία) γλώσσα. Συνεπώς, η πορεία από εδώ και πέρα καθορίζεται σχεδόν αποκλειστικά από το LLVM.

Πιο συγκεκριμένα, εάν ο χρήστης έχει ζητήσει να γίνουν βελτιστοποιήσεις, ο ενδιάμεσος κώδικας περνάει από το **opt**, ένα command line utility του LLVM το οποίο δέχεται ως είσοδο LLVM IR κώδικα και παράγει (πιθανώς) βελτιστοποιημένο LLVM IR κώδικα. Δεν θελήσαμε να κάνουμε υποχρεωτικό το στάδιο της βελτιστοποίησης του ενδιάμεσου κώδικα, ώστε να μπορούμε να πειραματιστούμε περισσότερο με τις διαφορές που προκύπτουν (με ή χωρίς βελτιστοποίηση) στη συνέχεια του pipeline, αλλά και για να μπορεί ο χρήστης να δει τον ενδιάμεσο κώδικα ακριβώς όπως τον παράγουμε εμείς μέσω των C++ bindings, χωρίς καμία αλλαγή (με χρήση της σημαίας -i, όπως περιγράφεται στην εκφώνηση της εργασίας).

Στη συνέχεια, το **llc**, ο compiler του LLVM, αναλαμβάνει να μετατρέψει τον ενδιάμεσο κώδικα σε object file, να το μεταγλωττίσει, δηλαδή, στη native assembly.

Τέλος, ο **clang** αναλαμβάνει να συνδέσει (link) το object file που έχει παραχθεί με τη standard βιβλιοθήκη της Alan (libalanstd.a), παράγοντας έτσι το τελικό native εκτελέσιμο πρόγραμμα.

2. Λεκτικός αναλυτής (lexer)

Αρχικά, θα

3. Συντακτικός αναλυτής (parser) και κατασκευή abstract syntax tree (AST)

Αρχικά, θα

4. Σημασιολογικός αναλυτής

Αρχικά, θα

5. Παραγωγή ενδιάμεσου κώδικα και βελτιστοποίηση

Αρχικά, θα

6. Παραγωγή τελικού κώδικα και βελτιστοποίησησ

Αρχικά, θα

7. alanc – ο τελικός μεταγλωττιστής

Σε αυτό το μέρος της παρούσας αναφοράς θα περιγράψουμε τον τρόπο χρήσης και λειτουργίας του τελικού εκτελέσιμου του μεταγλωττιστή, του **alanc**. Στο πρώτο μέρος παρουσιάσαμε συνοπτικά την γενική αρχιτεκτονική του script, ενώ σε αυτό το μέρος θα αναφερθούμε πιο αναλυτικά στην υλοποίησή του καθώς και στις σχεδιαστικές επιλογές που αποφασίσαμε να κάνουμε.

Αρχικά, παρατίθεται το μήνυμα που λαμβάνει ο χρήστης με την εκτέλεση της εντολής **./alanc -h**:

```
usage: alanc [-h] [-O] [-a] [-i] [-f] [-c] [-o OUTNAME] [infile]

ALANC - the Alan Limitless and Amazingly Neat Compiler

positional arguments:
  infile      (if -f or -i not given) the Alan source code to compile

optional arguments:
  -h, --help  show this help message and exit
  -O          optimize IR and final (assembly) code
  -a, --all   emit IR and assembly code in two seperate files (switched off by
              default)
  -i          read source code from stdin, print IR code to stdout, then exit
              (no executable is produced)
  -f          read source code from stdin, print final code to stdout, then
              exit (no executable is produced)
  -c          create object file and stop, skipping the linking phase
  -o OUTNAME  (if -f or -i not given) the name of the produced executable
              (default: a.out)
```

Όπως φαίνεται παραπάνω, υπάρχει μία σειρά από flags που προσφέρονται στον χρήση με σκοπό την παραμετροποίηση της λειτουργίας του μεταγλωττιστή.

Το flag **-O** ενεργοποιεί την βελτιστοποίηση τόσο του ενδιάμεσου (LLVM IR) όσο και του τελικού κώδικα (native assembly). Εξηγήσαμε στο πρώτο μέρος της αναφοράς την απόφασή μας να μην είναι υποχρεωτική η βελτιστοποίηση του ενδιάμεσου κώδικα.

Με χρήση του flag **-a (--all)** δημιουργούνται δύο επιπλέον αρχεία, ένα που περιέχει τον ενδιάμεσο και ένα τον τελικό κώδικα που παράγεται από τον μεταγλωττιστή, σε human-readable μορφή, κατά τις προδιαγραφές της εκφώνησης. Αποφασίσαμε να μην ορίσουμε ως default συμπεριφορά του μεταγλωττιστή την παραγωγή αυτών των δύο αρχείων (όπως φαίνεται, βέβαια, να υπονοείται από την εκφώνηση της εργασίας) ώστε να αποφύγουμε να δημιουργούνται δύο επιπλέον αρχεία κάθε φορά που ο χρήστης θέλει να μεταγλωττίσει ένα πρόγραμμα, τη στιγμή που απλά περιμένει να δημιουργηθεί ένα εκτελέσιμο. Με άλλα λόγια, δώσαμε σε αυτό το flag (και σε αυτήν την λειτουργικότητα του μεταγλωττιστή) μία διάσταση debugging.

Τα flags **-i** και **-f** ακολουθούν τις προδιαγραφές της εκφώνησης, όπως φαίνεται και στις περιγραφές τους στο μήνυμα παραπάνω. Αξίζει να σημειωθεί πως αποφασίσαμε να κάνουμε τη χρήση τους mutually exclusive, προς αποφυγή συγχύσεων. Επίσης, όπως φαίνεται και στα παραπάνω μηνύματα, αποφασίσαμε να μην παράγεται το τελικό εκτελέσιμο σε αυτές τις περιπτώσεις, δίνοντας την δυνατότητα στον μεταγλωττιστή μας να λειτουργήσει και ως μέρος κάποιου toolchain που θέλει να κατασκευάσει ο χρήστης, όπου και δεν περιμένει ή επιθυμεί τέτοιου είδους side effects.

Το flag **-c** έχει ακριβώς την ίδια σημασία με το αντίστοιχο flag του gcc: σταματάει στη φάση της μεταγλώττισης του πηγαίου προγράμματος, χωρίς να προχωρήσει στο στάδιο της σύνδεσης (linking) με τη standard βιβλιοθήκη της Alan και, τελικά, την παραγωγή του εκτελέσιμου προγράμματος. Τελικό προϊόν σε περίπτωση που δοθεί αυτό το flag είναι το αντίστοιχο object file, με κατάληξη “.o”. Σημειώνεται πως αυτό το αρχείο πάντα παράγεται, απλά σε περίπτωση που δεν

έχει δοθεί το εν λόγω flag, διαγράφεται μετά την παραγωγή του εκτελέσιμου προγράμματος.

Επιπλέον, δίνεται στο χρήστη η δυνατότητα να δώσει κάποιο όνομα που επιθυμεί στο τελικό εκτελέσιμο με τη χρήση του flag **-o** (όπως και στον gcc, το default όνομα για το παραγόμενο εκτελέσιμο πρόγραμμα είναι a.out).

Τέλος, ο χρήστης μπορεί να δώσει το όνομα ενός αρχείου προς μεταγλώττιση, ακριβώς όπως στον gcc και στον clang. Προφανώς, εάν δώσει κάποιο όνομα αρχείου, δεν μπορεί να έχει δώσει και κάποιο από τα flags **-i** ή **-f**.

Τα παραπάνω flags ορίστηκαν και υλοποιήθηκαν με βάση το module **argparse** της Python. Όταν το πρόγραμμα ξεκινά, γίνονται κάποιοι έλεγχοι σχετικοί με τα flags:

1. Ο χρήστης επιτρέπεται είτε να χρησιμοποιήσει ένα εκ των flags **-i** ή **-f** είτε να δώσει ένα όνομα αρχείου προς μεταγλώττιση.
2. Ένα από τα δύο παραπάνω πρέπει να έχουν δοθεί (δηλαδή, δεν επιτρέπεται να λείπουν και τα δύο)
3. Μόνο ένα εκ των flags **-i** ή **-f** επιτρέπεται να χρησιμοποιηθεί
4. Ο χρήστης επιτρέπεται είτε να χρησιμοποιήσει ένα εκ των flags **-i** ή **-f** είτε το flag **-a** (στα πλαίσια της σχεδιαστικής μας επιλογής να μην επιτρέπουμε side effects στην περίπτωση χρήσης των flags **-i** ή **-f**)
5. Δεδομένου του ότι η χρήση των flags **-i** ή **-f** συνεπάγεται μη παραγωγή τελικού εκτελέσιμου προγράμματος, δεν επιτρέπεται να χρησιμοποιούνται μαζί με το flag **-o**.

Σε περίπτωση που οποιαδήποτε από τις παραπάνω συνθήκες δεν πληρούται, η εκτέλεση του μεταγλωττιστή τερματίζεται με αντίστοιχο μήνυμα λάθους από το argparse.

Στη συνέχεια, ορίζεται το όνομα του προγράμματος (αποκόπτεται η κατάληξη εάν αυτή είναι “.alan”, αλλιώς μένει ως έχει), του object file που θα παραχθεί, καθώς και των αρχείων .imm και .asm κατά τις προδιαγραφές της εκφώνησης, σε περίπτωση που αυτά ζητήθηκαν από τον χρήστη.

Έπειτα, ορίζονται οι εντολές που εκτελούνται για τα 4 βήματα της μεταγλώττισης που ακολουθούν μαζί με τα αντίστοιχα flags τους. Τα 4 αυτά βήματα φαίνονται εποπτικά στο διάγραμμα του πρώτου μέρους της παρούσας αναφοράς:

1. Για την **παραγωγή του ενδιάμεσου κώδικα** χρησιμοποιείται το πρόγραμμα **alan** με μοναδικό όρισμα το όνομα του προγράμματος ώστε να χρησιμοποιηθεί εσωτερικά από το LLVM αλλά και για πιθανά μηνύματα λάθους. Δέχεται το πηγαίο κώδικα από το stdin και παράγει τον ενδιάμεσο κώδικα στο stdout.
2. Για την **βελτιστοποίηση του ενδιάμεσου κώδικα** χρησιμοποιείται το command line utility **opt** του LLVM, με το flag **-O3**, το οποίο καλύπτει τα ζητούμενα της πρώτης bonus μονάδας της εργασίας. Δέχεται τον ενδιάμεσο κώδικα από το stdin και παράγει (πιθανώς) βελτιστοποιημένο ενδιάμεσο κώδικα στο stdout. Αυτό το βήμα παραλείπεται σε περίπτωση που δεν έχει δοθεί από τον χρήστη το αντίστοιχο flag. Επιπλέον, μετά από αυτό το βήμα (είτε έχει εκτελεστεί είτε όχι) και εάν αυτό έχει ζητηθεί από τον χρήστη, τότε ο ενδιάμεσος κώδικας είτε αποθηκεύεται σε κατάλληλο .imm αρχείο και η εκτέλεση του μεταγλωττιστή συνεχίζεται είτε τυπώνεται στην οθόνη και το πρόγραμμα τερματίζει επιτυχώς.
3. Για την **παραγωγή του τελικού κώδικα** χρησιμοποιείται το command line utility **llc**, ο μεταγλωττιστής του LLVM, μαζί με το flag **-filetype=obj**. Αποφασίσαμε σε αυτό το στάδιο να παράγεται object file και όχι κάποια άλλη μορφή assembly ώστε να μπορούμε να υποστηρίξουμε ευκολότερα τη λειτουργία του flag **-c**, να είναι ευκολότερο το linking στη συνέχεια, αλλά και γιατί αποτελεί standard συμπεριφορά πολλών μεταγλωττιστών η παραγωγή ενός object file σε αυτή τη φάση, κάτι που μπορεί ο χρήστης να επιθυμεί. Σε περίπτωση που δόθηκε το flag βελτιστοποίησης, προστίθεται στα flags του llc το **-O3**, το οποίο καλύπτει τα ζητούμενα της δεύτερης bonus μονάδας της εργασίας. Η εντολή αυτή δέχεται τον (πιθανώς βελτιστοποιημένο) ενδιάμεσο κώδικα από το stdin και παράγει ένα

object file το οποίο αποθηκεύεται στο file system, στο τρέχον directory. Επιπλέον, μετά από αυτό το βήμα και εάν αυτό έχει ζητηθεί από τον χρήστη, τότε ο τελικός κώδικας είτε αποθηκεύεται σε κατάλληλο .asm αρχείο και η εκτέλεση του μεταγλωττιστή συνεχίζεται είτε τυπώνεται στην οθόνη, διαγράφεται το object file και το πρόγραμμα τερματίζει επιτυχώς. Για να συμβούν αυτά τα δύο εφόσον ζητηθούν, ξανακαλούμε τον llc, αλλά αυτή τη φορά χωρίς το flag **-filetype=obj**, ώστε να παραχθεί human-readable assembly αντί για ένα object file.

4. Το τελευταίο βήμα είναι **η σύνδεση του object file με τη standard βιβλιοθήκη της Alan (linking)**. Για αυτό το βήμα, εκτελείται απλώς η εντολή:
   `clang <objectfile> libalanstd.a –o <outname>`
   Μετά από αυτήν την εντολή, έχει παραχθεί το τελικό εκτελέσιμο πρόγραμμα με όνομα outname. Εάν έχουμε φτάσει σε αυτό το σημείο, διαγράφεται και το object file από το file system.

Σημειώνεται πως σε κάθε ένα από τα παραπάνω βήματα ελέγχεται στο stderr κάθε εντολής. Σε περίπτωση που δεν είναι κενό, το μήνυμα προωθείται στον χρήστη και ο μεταγλωττιστής τερματίζει, ανεπιτυχώς (με αρνητικό κωδικό εξόδου).

Για την εκτέλεση όλων των παραπάνω εντολών χρησιμοποιήθηκε το module **subprocess** ενώ για τον χειρισμό των directories, των paths και των αρχείων χρησιμοποιήθηκε το module **os**, και πάλι σε μία προσπάθεια για cross-platform υλοποίηση.