JANCY: The Return of the Pointers

Contents

[Introduction 1](#_Toc372545418)

[Inspiration 2](#_Toc372545419)

[C 2](#_Toc372545420)

[C++ 2](#_Toc372545421)

[Java/C# 3](#_Toc372545422)

[D 4](#_Toc372545423)

[Motivation 4](#_Toc372545424)

[Motivation №1, wheel-re-inventive 4](#_Toc372545425)

[Motivation №2, practical 5](#_Toc372545426)

[Motivation №3, philosophical 6](#_Toc372545427)

[Language design 6](#_Toc372545428)

[Design goals 6](#_Toc372545429)

[Compiler architecture 6](#_Toc372545430)

[Full feature list 7](#_Toc372545431)

[Full language grammar 9](#_Toc372545432)

[Current status & conclusion 9](#_Toc372545433)

# Introduction

What?! Yet another programming language? Well… yes. Meet Jancy. Before we even start, let me make a list of the features that might help you decide whether you are interested enough to keep reading.

* Object-oriented language with C-family syntax
* ABI (application-binary-interface) compatibility with C/C++
* Tailor-suited for being used as a scripting language from C/C++ host application
* Automatic memory management via accurate GC (garbage collection)
* Safe pointers and pointer arithmetic
* Properties (one of the most comprehensive implementations)
* Reactive programming paradigm (one of the first existing implementations in a general-purpose imperative language)
* Closures of functions and properties
* Weak pointers (do not retain the object)
* Multicasts and weak multicasts (do not require to unsubscribe)
* Explicit stack allocation, manual memory management, thread local storage
* RAII (resource-acquisition-is-initialization) paradigm support
* Const-correctness support
* Multiple inheritance support
* Perl-style formatting support
* …and many more – no, seriously! ;)

Now let’s slow down a bit and take a look back.

# Inspiration

### C

A long time ago, in a galaxy far, far away… and to be more precise, in the end of sixties (starting in 1969) at AT&T Bell Labs Dennis Ritchie has created the most important programming language of all times: the C language. It’s difficult to overestimate its influence on the development of computer science. Even now, over 40 years since its birth, good old C is still among the most used programming languages – regardless of which exact rating of programming language popularity one would choose. Moreover, the top of the list would always be shared with its direct descendant C++ and other sibling languages whose syntax was derived from or heavily influenced by C: Java, JavaScript, C#, Objective C. And if the rating would only consider the usage in systems programming, then the gap between C & C++ and the nearest competitor becomes simply embarrassing.

I love C. Syntax of C is simple, elegant, expressive and universal. And what’s very important in systems programming, the layer of code generated by the compiler is thin and predictable: an experienced programmer could take a look at the C snippet and immediately estimate how the same snippet would look in assembly language after compilation.

### C++

Direct descendant of C which emerged more than a decade later in 1983, C++, unfortunately cannot say the same about itself. C++ has passionate supporters and no-less passionate and irreconcilable opponents. Holy wars on C++ go on all the time, on any programming forum, any weather, any season. The arguments usually go around the following statements.

* C++ is indeed extremely expressive and powerful
* High performance of C++ programs

Actually, the second in no small measure is the consequence of the first. Thanks to the expressive power of C++ programmers are able to create high-performance code.

* Difficulty of C++

Language is so difficult that it earned itself a place in programmers’ jokes – and rightfully so. Together with excellent expressive power come numerous and diverse capabilities to “shoot your own foot”. It’s so much easier to write bad C++ than it is to write bad C!

The layer of compiler-generated code has significantly grown in size and complexity and became much less predictable even by very experienced programmers.

It’s quite possible that largely because of the above even after good C++ compilers came into mainstream, good old C remained the main language for systems programming. Windows, Linux, Doom and Quake are all written in plain C while C++ was around for many, many years.

* C++ lacks many capabilities which are expected to be built-in into the modern general-purpose language, such as automatic memory management or reflection.

As a bottom line, it’s totally possible to efficiently use C++ as a programming language for building software of any level of complexity. However, a very high qualification of the developer is an absolute must. Moreover, even a qualified specialist from time to time shoots his own foot or feet of his colleagues.

### Java/C#

Managed languages came to finally protect the feet of developers. Yet another decade past creation of C++, Java emerged in 1995. Microsoft-created managed platform .NET with its flagman language C# initially looked like a clone of Java, thoughtlessly copying both good and bad but was later quite ingeniously extended. In the following paragraphs though, I will write about Java only. But most of it also applies to C#.

The Java compiler doesn’t generate the machine code for the specific processor, but Java byte-code, which is then interpreted or JIT-compiled by the Java Virtual Machine at run-time. This allows reusing the same binary module on multiple architectures without re-compiling.

The Java Virtual Machine also provides a built-in facility for automatic memory management via garbage collection. Java developers do not need to worry about explicit freeing of memory blocks allocated for various objects since the garbage collector will take care of that. This, of course, does not mean that memory leaks in Java are non-existent. Besides, there are situations where it’s highly desired to manage memory manually instead of relying on the convenient but not-so-deterministic facility of garbage collection. But generally speaking, built-in accurate garbage collection is one big fat plus, which is admitted by most (adequate) C/C++ proponents:

* A large part of any C/C++ program is dedicated to freeing memory. For example, if memory management was automatic, most (not all, of course – but absolute most!) C++ destructors would simply become irrelevant cause all they do is free memory
* The alternatives of automatic memory management available to C++ developers (namely, variations of reference counting or conservative garbage collection) are in many respects inferior to built-in accurate GC.
* Passing objects between C++ modules is largely handicapped by the lack of standardized facility of automatic memory management
* The whole class of bugs related to premature freeing or the opposite, non-freeing of memory blocks – simply goes away

As a consequence of the above, the built-in garbage collection simplifies and the therefore speeds up and reduces the cost of the development cycle.

Java syntax resembles heavily simplified C++. As a result and also as a good evidence of this simplification Java grammar could be expressed with context-free LR (1) – unlike C++ which has context-sensitive grammar which is widely considered one of if not the most complicated grammars in modern programming languages. This simplification is of course a big advantage for everybody, from regular Java developers and to developers of code-assisting IDEs. Simple is good!

Besides grammar simplification, the following language facilities were also given the boot.

* Namespaces
* Type aliases (typedefs)
* Global functions and variables
* Pointers and pointer arithmetic
* Structs and unions
* Explicit stack allocation, RAII and manual memory management
* Operator overloading
* Multiple inheritance
* Const-correctness
* Default argument values

While I have no intention of starting yet another holy war, I will risk saying that most of these terminations were unjust. Of course there was reasoning for the Java designers’ part. Like simplification of syntax, compiler and virtual machine; or like improving language safety via not giving a developer potentially dangerous tools and so forth. But it’s so easy to give examples where each of the above could be very helpful to make the code sometimes more elegant, sometimes more logical, and often times more efficient!

All in all the built-in accurate garbage collector, familiar and simple C-like syntax as well as the safety and ability to run the same binary on any platform given the fact it has Java Virtual Machine led to unquestionably huge success of Java. If we take a look at ratings of the programming language popularity (which I’ve started this article with) we will see that Java not only consistently remains among the top positions but also slowly but steadily improves its ratings year by year.

### D

Although the D programming language often doesn’t even make it into top ten or top twenty of the above ratings, it would be unfair not to say anything about this interesting language. The D programming language was created by Walter Bright as a re-thinking of C++ at 1999 – about the same time Microsoft released the first version of C#. The main motto of the D language design is combining the performance of C/C++ with the convenience and safety of the managed world.

D has full ABI (application binary interface) compatibility with C/C++ programs, supports pointers to data and functions, manual memory management, allows mixing in portions of assembly language and at the same time provides facilities from the managed world such as garbage collection, nested functions and closures, strong meta-programming facilities and more.

D is an awesome language and gave me tons of inspiration. Years back I was seriously considering and was running experiments on using D as a scripting language in our products, but the early stages D at that time, and most importantly the absense (again, at that time) of a back-end available for commercial usage put that on hold.

Besides, D has its share of rather questionable design decisions (for example, scope guards or auto-conversion of functions into properties). Also, unsafe pointers and the possibility of mixing in assembly snippets largely denies the usage of D as a safe scripting language.

# Motivation

### Motivation №1, wheel re-inventive

One of the remarkable qualities of the Russian education was, and I hope still is, the nurturing of a critical view on things. Starting with senior high and certainly in the university we were taught not to blindly accept what is written in the textbooks or what is told by a lecturer at the blackboard. I remember at the very first lecture on physics, our teacher without breaking a straight face has given us a proof that 2 equals 5 (these are the designation marks in the Russian education system by the way), and right after that followed the proof that the acute angle equals the obtuse one. Then he put his idea into words: try not to thoughtlessly accept and memorize something told or shown by undoubtedly smart people (I suspect he meant himself) but rather comprehend and understand why it is the way it is, why it works like this, and not like that. Perhaps, you will find an error in their reasoning. Perhaps, you will find another way to do the same which would be somehow better than the original one. Or perhaps, quite the opposite – you will realize that it’s better to apply your time and efforts to something else instead of trying to fix something that works already.

On the one hand, I am extremely grateful for being taught and trained to use this over-analytical way of thinking. On the other hand, to a certain extent I have become an irredeemable wheel inventor – and that is of course not the title to be proud of. To my defense I can say though that another rule taught in university – namely, the rule to study existing types of wheels before starting to build your own one – without a doubt reduced the number of wheels I have created throughout my life by a magnitude. Maybe even two.

I was programming for the last 20 years, 15 years professionally. In every language I happened to use at work or in my leisure programming, my internal wheel inventor was constantly looking for and finding things to improve. It so happened that my area of professional expertise is systems programming, particularly compilers, debuggers and other development tools. It is therefore not surprising that for many years I have had little toy languages as proving grounds for experiments. Taking into account my confession in love to the C language, it’s not surprising that these languages always had a C-like syntax.

Until recently though, these languages had no name – despite the fact that previous versions (or should I say incarnations? Because most of the time the new version did not have a single line from the previous one) were used in commercial products as DSLs (domain-specific scripting languages). At one point I decided to put a stop to this anonymity and after a rather brief (lucky!) linguistic odyssey I found the name: Jancy.

It’s easy to see that Jancy is an acronym: [in between] JAva aNd C. One could tell that I took the Java as a starting point and brought in the features and facilities I love in C/C++. Or the other way around: that it’s simplified C/C++ which is taught the Force of the managed world. In that respect Jancy targets similar design goals as D. Plus, of course, I implemented the features which are not found anywhere else (ok, maybe they could be found, but they certainly have not made their way to the mainstream programming languages yet) – the features that I always dreamed of and which passed the test of my toy languages proving grounds.

So… there, I confess. One of the main motivations of creating Jancy was feeding my internal wheel inventor.

### Motivation №2, practical

But there was also a practical motivation for turning leisure development into a full-scale project. A couple of years ago our company released a product named I/O Ninja, one of the modules of which had to analyze binary data blocks and write analysis results into a log file. To avoid accusation in advertisement, I am purposefully not going to dive into details or even explanations what this product was all about and why it was needed. It was necessary to analyze and generate binary data blocks – that’s it.

Since we wanted to make analyzing as flexible as possible, we had to move the logic of analyzing into plug-ins or scripts. And since in the long run we planned to give our users possibility to add new analyzers or modify existing ones – scripts seemed like a better approach (plug-ins would mean the necessity of publishing the SDK; our users would also need to possess both C/C++ development tools and the required qualification to undertake C/C++ development).

So, script it is. But herein lies the problem: it’s difficult to perform analyzing/generating binary data blocks without data pointers and pointer arithmetic. Without these facilities it’s all about taking the bytes at fixed indexes and clueing them into words, and this generates unreadable and unsupportable code. Working with binary data is best done with structs and pointer arithmetic. Period. There are no well-known/recognized scripting languages with data pointer supports suitable for embedding into C/C++ host applications – not back then, not now. And even if there are some experimental languages Google knows nothing about, they are at no advantage over an in-house-invented wheel.

The first version of I/O Ninja was actually built on a plug-in architecture; the second one – on scripts written in a Jancy prototype (unsafe pointers, automatic memory management based on reference-counting, rudimentary class support and a hand-made virtual machine). The third version is under active development right now and based on scripts in the full Jancy language.

This product is a practical motivator for creating Jancy and will serve as an excellent real-life proving ground.

### Motivation №3, philosophical

Let me philosophize just a bit more before getting to Jancy’s design goals and feature list. In no way I want to say that the creation of Jancy was “forced”, that the problem could not have been solved with different languages, different tools, different approaches. It certainly could have been. There always is another way.

But that’s the story of progress: most innovations simply make some process a little bit better: a little bit faster, or easier, or more efficient. It was totally possible to write everything in good old C. And even directly in assembly language! Just like it’s still totally possible to move from point A to point B on horses or by walk. Or use candles to illuminate the house. But any process can always be improved a little bit. And then improved a bit more. These little tiny bit of improvements, each of which is by no means revolutionary (hey, it was totally possible to achieve the same result before!) – add up to bigger ones, and even bigger ones. And finally they add up to the extent that they allow doing something which was simply impossible earlier.

These little non-revolutionary improvements is what we are trying to introduce.

# Language design

### Design goals

The main model of usage considered while designing Jancy, was the usage as a scripting engine embedded into the host C/C++ application. In the long run, of course, it would be flattering to see Jancy used as a general-purpose language tailored for systems programming.

The main design goals of Jancy are:

* Usage as a scripting language from the host C/C++ application
* Familiar C-like imperative syntax
* Safe pointers and pointer arithmetic
* Automatic memory management via accurate garbage collection

### Compiler architecture

The Jancy compiler uses a lexical analyzer generated by Ragel (a universal finite state machine compiler). It is perfectly suited for building lexers thanks to the convenience and expressiveness of its input language and even more importantly, to the unbeatable performance of the output code.

As a backend Jancy uses LLVM. It allows making use of a reliable optimizer and native code generator for a wide variety of platforms at once. It also significantly simplifies ensuring of ABI compatibility with C/C++. All in all, LLVM usage as opposed to hand-written code generator or virtual machine was not even questioned – ever since I first laid my hands on LLVM.

As for the syntax analyzer though I couldn’t resist to create my own ~~wheel~~ generator of table-driven top-down LL (k) parsers. What’s wrong with ANTLR, Coco, Yacc/Bison, Lemon and other undoubtedly respectable and tried-and-true parser generators is an interesting topic – but for a separate discussion. What’s important here is that the Jancy compiler doesn’t use a hand-written recursive descent but a generated parser. Among other things it also means that we have permanently relevant BNF grammar which can be, let’s say, printed out, discussed and easily adjusted. Grammar of Jancy belongs to context-sensitive LL (2).

The output of the parser is LLVM IR without intermediate generation of an AST: unlike in bottom-up parsers, in top-down parsers it is rather convenient to perform semantic analysis and generate IR parallel to parsing, right as the parser matches the grammar rules.

The model of collaboration between parser and lexer is peeped in Lemon. The Jancy parser does not call the lexer. Instead, the external loop asks the lexer to tokenize the next chunk of source and feeds tokens to the table-driven parser. This model allows incremental parsing of incomplete compilation units (supports pause and resume so to speak), plus, like in any table-driven parsers, the memory for rule attributes is allocated not on the stack (which requires artificial nesting level limitations due to the danger of a stack overflow) but on the heap.

Syntactic/semantic analysis is performed in multiple passes (mostly two, some rules require three). This, however, does not mean re-tokenizing the source multiple times. The second pass is needed to allow the usage of global entities (namespaces, types, variables, functions and properties) before their declaration which could happen to be below or even in a separate compilation unit. The third pass is necessary for preliminary calculation of reactor class layouts.

### Full feature list

Jancy has a long list of features, improvements and innovations of varying levels of importance.

* Full ABI-compatibility with C/C++
  + After proper declaration of data types in Jancy scripts and in the host C/C++ application it’s possible to directly pass data through arguments and return values without the need to explicitly push and pop the stack of the virtual machine, packing/unpacking them into variant-like containers etc. All the following types are supported:
    - All primitive C/C++ types (also integer types with inversed byte order)
    - Structs (with arbitrary pack factor)
    - Unions
    - Arrays
    - C/C++ pointers to data or to functions
* Classes
  + Classes are special types of data with ancillary header
  + All the standard features of classes are supported:
    - Constructors / destructors (both non-static and static)
    - In-place field initialization
    - Operator overloading
    - Abstract and virtual functions
  + Simple multiple inheritance (no infamous C++ virtual inheritance)
  + Implementation of methods could be in-class, or out-of-class (like in C++).
  + Support for “preconstructor”s – special methods that should be called before any of the overloaded constructors: analogue of Java initializer blocks.
* RAII and full control over data storage
  + Support for explicit stack allocation and RAII paradigm (like in C++)
  + Support for allocating memory for multiple objects at once (by declaring class field members like in C++), without the need of explicitly calling “new” for every object
  + Built-in support for TLS variables (“thread” storage specifier)
  + Manual memory management of unmanaged heap with “heapu new” / “delete”
  + Storage specifiers: “static”, “heap”, “heapu”, “stack”, “thread“ control which memory should be used for particular data entities (field, variable, “new” operator)
* Pointers to data
  + Const-correctness
  + Safe pointer arithmetic
  + Support of weak class pointers (the ones that do not retain the object)
* Properties
  + One of the most comprehensive implementations of properties
  + Natural form of simple property declarations
  + Full form for declaring properties of arbitrary complexity (with overloaded setters, member fields, ancillary methods etc)
  + Indexed properties (properties that expose semantics of arrays)
  + Autoget-properties eliminate the routine of writing trivial getters
  + Bindable-properties notify subscribers the moment they change
  + Bindable-data automatically generate bindable properties with trivial getters and setters
* Pointers to functions and properties
  + Function and property pointers can capture the value of arguments
  + Automatic creation of thunk functions when argument signatures do not match but could be converted
  + Support of weak pointers to functions and properties
  + Schedule operator on a function pointer creates an ancillary function pointer which ensures execution of the original one in the desired context (in the proper thread, with mutex held, etc)
* Multicasts
  + Multicasts allow accumulation of function pointers and then calling them all at once
  + Event pointers to multicasts only allow subscribing/unsubscribing but not calling
  + Two kinds of conversions from a multicast to a function pointer: live & snapshot
  + Support of weak multicasts (the ones that do not need explicit unsubscribe)
* Reactive programming
  + One of the first implementations of reactive programming paradigm in an imperative language
  + The dilemma of mixing reactive paradigm with imperative data flow is solved using dedicated blocks of reactive code – so-called “reactors”
  + Expressions within reactors will be automatically re-evaluated upon the change of bindable r-values
  + Syntax constructs for assigning blocks of code as on-change handlers without the need of explicit subscribe-unsubscribe
* Exception handling
  + Exception model of Jancy is merely syntactic sugar over the old C-style error code model
  + As a result, it is completely transparent and compatible with any external language
  + “catch” and “finally” can be declared in any scope
  + When calling the same function, the developer can use either error code check, or exception semantics – depending on what’s more appropriate or convenient in this particular case
* Dual access control model
  + Jancy only has 2 access specifiers: public and protected
  + Each namespace splits the rest of the namespaces into 2 teams: “friends” and “aliens”. Friends can have access to anything, aliens – to “public” items only
  + Dual modifier “constd” for declaring type of data which looks like “const” for aliens, and “non-const” for “friends”. Designated to eliminate the routine of creating trivial properties whose only purpose is providing read-only access
  + Dual modifier “eventd” for declaring event pointers. Aliens can only subscribe/unsubscribe (“event” semantics) while friends are also able to fire the event (“multicast” semantics)
* Special literals
  + Hex-literals for convenient declaration of constant binary data blocks
  + Formatting literals dynamically generate strings substituting the values of expressions referenced from within the literal (perl-style formatting)
* Miscellaneous
  + “once”/thread “once”
  + Comprehensive implementation of curly-intitializers
  + “break-n”/”continue-n”: controlling outer loops
  + “basetype-n”: convenient referencing of base types
  + “enumf”: flag enumerations
  + “extend”: extending functionality of existing classes with methods
  + “alias”: expression aliases
  + Scopes-in-switch
  + Constructors/destructors of compilation units

### Full language grammar

axl\_jnc\_Lexer.rl

axl\_jnc\_Decl.llk

axl\_jnc\_DeclarationSpecifier.llk

axl\_jnc\_Declarator.llk

axl\_jnc\_Expr.llk

axl\_jnc\_Expr\_s.llk

axl\_jnc\_NamedTypeSpecifier.llk

axl\_jnc\_Parser.llk

axl\_jnc\_Stmt.llk

axl\_jnc\_Stmt\_0.llk

# Current status & conclusion

Jancy is a cross-platform project that targets Windows, Linux as well as Mac (planned) and will generate native code for any architecture supported by LLVM. Jancy is not currently open-source, but may become so in the future. A Jancy plugin for NetBeans with code assist support and debugging capabilities via GDB is under active development and is likely to be released together with the new Jancy-based version of I/O Ninja.

At present time we have a working demo of compiler available for testing online, without any download or/and installation. Certain features which could be expected from a modern language (such as reflection, generics/templates, lambda-functions etc) are not implemented yet, but they certainly will be there in one of the future releases. At the same time, the features that are implemented and thought to be ready, might not work quite as expected after being exposed to rigorous testing through the web – it’s a first public release after all. Bug reports are of course more than welcome. We believe that during development of Jancy-based I/O Ninja and with the help of web-based testing we will be able to polish both the language and the compiler and make it suitable for use in a wide class of applications.

We certainly hope that we have managed to spark your interest and would be happy to receive bug reports, questions and general feedback from the community of developers.