# Ecstasy: XVM and the XTC Language

Version: DRAFT-20180730

*Software is music: Elegance in simplicity; harmony in vision; rhythm in motion.*

# Introduction

Imagine a system of execution. The ingredients are well known at a conceptual level: Instructions, conditionals, control flow, state, loads, and stores. Processors, virtual machines, languages: Each a tool to support the description of a desired outcome. Processes, threads, fibers, tasks, jobs: Each a picture of execution.

We create systems of executions as abstractions that exist between the definite world of digital processing and the indefinite and limitless world of programming. In the immortal words of Fred Brooks writing in *The Mythical Man Month*:

Finally, there is the delight of working in such a tractable medium. The programmer, like the poet, works only slightly removed from pure thought-stuff. He builds his castles in the air, from air, creating by exertion of the imagination. Few media of creation are so flexible, so easy to polish and rework, so readily capable of realizing grand conceptual structures.

Yet the program construct, unlike the poet's words, is real in the sense that it moves and works, producing visible outputs separate from the construct itself. It prints results, draws pictures, produces sounds, moves arms. The magic of myth and legend has come true in our time. One types the correct incantation on a keyboard, and a display screen comes to life, showing things that never were nor could be.

This document describes a system of execution, the XVM[[1]](#footnote-1). It exists in the abstract as defined by this document, and much like similar technologies that have preceded it, it is expected that compliant implementations could be implemented in software, with execution efficiency provided by translation to the execution language of the underlying machine, whether it be hardware or software.

It is this simultaneous duality of purpose – to provide both a natural and immersive abstraction, and to maximize the potential efficiency of execution – that creates a dramatic tension in design. Most computer languages and their associated execution machinery are examples of the necessary trade-offs made to resolve this tension; very few examples exist without evident compromise. At one end of the spectrum are the hardware-centric models, starting with machine code, then assembly code, and progressing into the realm of increasing abstraction with languages such as FORTRAN, COBOL, and C; these programming languages allow the expression of intention using constructs that translate very directly to the mechanics of the underlying machine, thus simplifying the potential for optimization of execution in terms of both time and space. As one proceeds along the language spectrum, abstractions emerge that are less and less a reflection of the underlying machinery, and more and more the attempt to realize concepts of ideal algorithmic and structural composition. These abstraction-centric models include object-oriented languages such as Smalltalk, Java, and C#, and functional languages such as Haskell, Erlang, and Clojure; these programming languages provide powerful abstractions that allow designs and concepts to translate very directly to the mechanics of the language, optimizing for some combination of simplicity, brevity, readability, consistency, conceptual purity, predictability of execution, code re-usability, portability, speed of development, and so on.

There is nothing more crucial in understanding a system of execution than understanding how the designers of the system chose to resolve this natural tension. The two questions that must always be answered are these: (i) How does the system propose to map its capabilities to the hardware instruction set, the hardware memory model, other hardware capabilities, and the operating system(s) within which it will execute; and (ii) how does the system propose to represent its capabilities to the software developer?

To that end, the following observations and analyses are made, in an attempt to capture both the purpose and the rationale behind the decisions that are made to resolve this fundamental tension in this particular system of execution.

## On Priorities

All designs have priorities, but only some designs begin with the end in mind. When priorities are not explicitly stated, it is easy to chase the priorities that most effectively combine compulsive emotional appeal with apparent ease of implementation, in lieu of the priorities that are most valuable to the intended audience of a design. In order to create a coherent design that serves the intended audience, this specification began with a conscious discussion about priorities, and a series of explicit decisions as to what those priorities would be:

1. Correctness, aka Predictability. The behavior of a language must be obvious, correct, and predictable. This also incorporates *The Principle of Least Surprise*.
2. Security. While generally not a priority for language design, it is self-evident that security is not something that one *adds to* a system; security is either in the foundation and the substrate of a system, or it does not exist. Specifically, *a language must not make possible the access to (or even make detectable the existence of) any resource that is not explicitly granted to the running software*.
3. Composability. High-level computer languages are about composition. Specifically, *a language should enable a developer to locate each piece of design and logic in its one best and natural place*.
4. Readability. Code is written once, and referenced many times. What we call “code” should be a thing of beauty, where *form follows function*.
5. Lastly, a language must be recursive *in its design*. There is no other mechanism that predictably folds in complexity and naturally enables encapsulation. *It’s turtles, the whole way down*.

## On Hierarchical Organization

Many a software developer has referenced this saying:

*When the only tool that you have is a hammer, every problem begins to look like a nail.*

That is not to imply that a hammer is not useful. If there is one conceptual hammer that – more so than any other – has repeatedly proven its merit for managing – and *hiding* – complexity, that would be the concept of *hierarchy*. File systems are hierarchies. B\*Trees and binary trees and Patricia tries and parse trees are hierarchies. Most documents are internally organized as hierarchies, including the common HTML, XML, and JSON formats. Most graphical user interfaces are modeled as hierarchies. Many programming languages leverage hierarchy to provide nesting, information hiding, scoping, and identity. How is it that such a simple concept can be so universally useful?

First of all, hierarchical organization enables very simple *navigation*. What this means is that at any arbitrary point – called a *node* – in a hierarchy, there is a well-known set of operations that are possible, such as navigating from the current node to its *parent* node, and navigating from the current node to any of its *child* nodes. If a node does not have a parent, then it is the *root* node, and if a node does not have any child nodes, then it is a *leaf* node.

Child nodes are *contained* within their parent node. Each child is uniquely identifiable by its parent, for example by a name or some other unique attribute. A hierarchy is *recursive*; at any point in the hierarchy, from that point down is itself a hierarchy. Since a hierarchy has a single root and is recursive, each node in the hierarchy is uniquely identifiable in the hierarchy by combining the identities of each successive node starting with the root and proceeding down to the node; this identity is the *absolute path* to that node. It is possible to navigate between any two nodes in the same hierarchy by combining zero or more child-to-parent navigations with zero or more uniquely identifiable parent-to-child navigations; the sequence of these steps is a *relative path* between two nodes.

These basic attributes of a hierarchy combine in amazingly powerful ways. For example, since each node is itself the beginning of a hierarchy of any size, it is possible to refer to that entire sub-hierarchy simply by referring to that one particular node; this effectively *hides* the recursive complexity contained – or *nested* – within that node. As a result, it is possible to add, copy, move, or remove a hierarchy of any size simply by adding, copying, moving, or removing the node that is the “root” of that hierarchy.

Using a hierarchy, it is incredibly simple to construct the concept of *scope*. For example, a scope could include only a specific node, or it could include a specific node and all of its child nodes recursively to its *descendent* leaf nodes, or it could include a specific node and its *ancestor* nodes to the root node, or any other combination of inclusion and exclusion that could be described in an unambiguous manner.

These concepts are incredibly simple, yet at the same time incredibly powerful, and are leveraged liberally throughout the XVM, from managing and hiding complexity for developers, to managing memory in an execution context.

## On Predictability vs. Performance

In the course of researching language design, one preterdominant concern emerges: That of execution performance. While many different concerns are evaluated and balanced against each other in a typical design, and while the goal of performance is often explicitly ranked in a position of lesser importance, in reality there is no one goal more considered, discussed, and pursued. Regardless of the importance assigned to performance as a language goal, performance to the language designer is a flame to a moth.

Perhaps it is simply best to admit, up front, that no language can survive – let alone succeed – without amazing feats of performance. Yet performance as a goal tends to conflict with other goals, particularly with respect to manageability, serviceability, and the quality of the abstractions that are presented to the developer.

Beginning programmers often ask: “Is language *A* faster than *B*?” After all, no one wants to be using a slow language, any more than someone would want to buy a slow automobile or a slow computer. Speed is a thrill, and speed in software execution holds no less attraction than a souped-up hot rod with a throaty growl and a body-rumbling ride.

The corollary to that question is “Why is language *B* slower than *A*?” The answers to this question tend to be very illuminating. Take any two languages that compile to and execute as native machine code, and compare their performance for a given task. Despite running on the same hardware, and using the same hardware instruction set, one may be dramatically slower than the other, by a factor of 10%, or 100% (half as fast), or even 1000% (an order of magnitude slower). How is such a difference possible?

The answer lies in translation, specifically in the automated translation of idioms. A language such as C selected idioms that closely represented the underlying machine instructions of the day, which allowed programmers to write simple programs that could be almost *transliterated* from C to machine code. In other words, the language chose as its abstractions the same set of abstractions that the CPU designers were using, or abstractions that were at most one translation removed from the machine’s abstractions. This allowed for very simple compilers, and made it extremely simple to support localized optimization.

A localized optimization is an optimization in the compiled code that can be made using only information about code that is *most local to* the code that is being optimized; in other words, information outside of the scope of the small amount of code being optimized is not even considered. Many optimizations in the C language, for example, are extremely local, such that they can be performed without any information about code outside of a particular expression or statement; it is hard to imagine more localized optimizations.

However, there is a trade-off implicit in achieving such simple and direct optimizations: The abstractions provided by the language are constrained by the abstractions provided by the CPU. As one could rightfully surmise, hardware abstractions tend not to be very abstract, and the abstractions provided by hardware instruction sets tend to be only slightly better. In its early days, the C language was jokingly referred to as “assembly with macros”, because as a language, it was only slightly higher level than assembly itself.

Computing efficiency is often stated in terms of a tradeoff between time (CPU) and space (memory); one can often utilize one in order to optimize for the other, subject to the law of diminishing returns. Unfortunately, there is no single simple metric that captures computing efficiency, but if the trade-off between time and space appeared as a graph with time on one axis and space on the other, it might generally resemble the shape of the curve *y=1/x*, which most closely approaches the origin at (1,1). If there were a single computing efficiency measurement for a programming language, it could arguably be represented by the closest that this trade-off curve approaches the origin (0,0), which distance could be considered the minimum weighted resource cost. To calculate a language’s efficiency for a particular purpose, one would calculate the *inverse* of the minimum weighted resource cost.

While one can easily speak about efficiency in hypothetical terms, a benchmark is a wonderful servant but a terrible master. The path chosen in the design of the XVM is to consciously avoid limits on potential efficiency by consciously avoiding contracts whose costs are not clearly defined and understood. This approach can be explained in the inverse, by way of example in existing languages and systems: Often, features and capabilities that were considered to be “easy” implementation-wise and “free” efficiency-wise when they were introduced, ultimately emerged as the nemesis to efficiency, due to the inflexibility of the programming contracts that these features and capabilities introduced[[2]](#footnote-2).

To understand this, it is important to think of abstractions as a two-sided coin: On one side, we see the benefit of the abstraction, which allows a programmer to work with ever-larger and more powerful building blocks, while the other side of the coin represents the cost of the abstraction, which is called the *contract*. Imagine, for example, having to build something in the physical world, out of actual matter. One could conceivably build a structure out of atoms themselves, assembling the necessary molecules and arranging them carefully into perfectly formed crystalline structures. The contracts of the various elements are fairly well understood, but yet we wouldn’t try to build a refrigerator of out individual atoms.

One could imagine that building from individual atoms is the equivalent of building software from individual machine instructions, in two different ways: First, that the refrigerator is composed of atoms, just like all software is executed at some level as machine instructions; and second, that as the size and complexity of the constituent components increase, the minutiae of the contracts of the sub-components must not be surfaced in the contracts of the resulting components – those details must be hidable! This purposeful prevention of the surfacing of minutiae is called *encapsulation*, and exists as one of the cornerstones of software design. It is why one can use a refrigerator without knowing the number of turns of wire in the cooling pump’s motor, and why one can use a software library without worrying about its impact on the FLAGS register of the CPU.

Ultimately, it is the recursive composition of software that creates challenges for optimization. While low level optimizations are focused on the creation of more efficient low level code, higher level optimizations rely on explicit knowledge of what portions of a component’s contract – or the contracts of the various sub-components – can be safely ignored. In other words, the optimizer must be able to identify which contractual effects are ignored or discarded by the programmer, and then leverage that information to find alternative execution solutions whose contracts manage to cover at least the non-discarded and non-ignored contract requirements. Higher-level optimizations target the elimination of entire aspects of carefully defined behavioral contracts, and as a result, they typically require extensive information from across the entire software system; in other words, high-level optimizations tend to be non-localized to the extreme! No software has been more instrumental in illustrating this concept than Java’s Hotspot virtual machine, whose capabilities include the inlining of potentially polymorphic code by the determination that the potential for dynamic dispatch is precluded, and the elimination of specific memory barriers in multi-threaded programs as the result of escape analysis.

To enable these types of future optimizations, the contracts of the system’s building blocks must be explicit, predictable, and purposefully constrained, which is what was meant by the goal of “consciously avoiding contracts whose costs are not clearly defined and understood.” The contracts in the small must be encapsulatable in the large, which is to say that contracts must be composable in such a way that side-effects are not inadvertently exposed. It has been posited[[3]](#footnote-3) that “all non-trivial abstractions, to some degree, are leaky,” but each such leak is eventually and necessarily realized as a limit to systemic efficiency.

## On God, Turtles, Balloons, and Sandboxes

Wikipedia defines a software sandbox as follows[[4]](#footnote-4):

In computer security, a sandbox is a security mechanism for separating running programs. It is often used to execute untested or untrusted programs or code, possibly from unverified or untrusted third parties, suppliers, users or websites, without risking harm to the host machine or operating system. A sandbox typically provides a tightly controlled set of resources for guest programs to run in, such as scratch space on disk and memory. Network access, the ability to inspect the host system or read from input devices are usually disallowed or heavily restricted.

In the sense of providing a highly controlled environment, sandboxes may be seen as a specific example of virtualization. Sandboxing is frequently used to test unverified programs that may contain a virus or other malicious code, without allowing the software to harm the host device.

In the physical world, in which children play with sand, there are two common styles of sandbox. The first is constructed from four equally sized wooden planks, each stood on its long edge to form a square box, fastened in the corners, and then filled with sand. The second style is typified by a large green plastic turtle, whose “turtle shell” is the removable top that keeps the rain out, and whose “body” is the hollow bowl that keeps the sand in. Both styles hold sand and allow a child to dig tunnels and build sand-castles, but there is one major difference: When a child tunnels too deeply in the wooden-sided sandbox, the tunnel burrows past the sand and into the soil beneath, while the tunnel depth in the turtle sandbox is strictly limited by the plastic bowl.

Software sandboxes tend to mirror these physical types, in that *the dirt often lies beneath*. In other words, the sandbox attempts to protect the resources of the system, but a determined programmer will eventually be able to find a way through. The only way that a language runtime as a sandbox can ensure the protection of the underlying resources of a system is for the sandbox itself to completely lack the ability to access those resources. Thus, the purpose of the sandbox is to defend against privilege escalation:

Privilege escalation is the act of exploiting a bug, design flaw or configuration oversight in an operating system or software application to gain elevated access to resources that are normally protected from an application or user. The result is that an application with more privileges than intended by the application developer or system administrator can perform unauthorized actions[[5]](#footnote-5).

As a language runtime designer, it is not sufficient to simply distrust the application code itself; one must distrust the entire graph of code that is reachable by the application code, including all third party libraries, including the language's own published runtime libraries, and including any internal libraries that come with the runtime that are accessible. Or, put another way, if there is a possible attack vector that is reachable, it will eventually be exploited. To truly seal the bottom of the sandbox, it is necessary to disallow resource access *through* the sandbox altogether, and to enforce that limit via transitive closure.

But what good is a language that lacks the ability to work with disks, file systems, networks, and network services? Such a language would be fairly worthless. Ecstasy addresses this requirement by employing *dependency injection*, which is a form of *inversion of control*. To comprehend this, it is important to imagine the container functionality not as a sandbox, but as a balloon, and our own universe as the primordial example.

Like an inflated balloon, the universe defines both a boundary and a set of contents. The boundary is defined not so much by a location, but rather by its impermeability – much like the bottom of the green plastic turtle sandbox. In other words, the content of the universe is fixed[[6]](#footnote-6), and nothing from within can escape, and nothing from without can enter. From the point of view of someone within our actual universe, such as you the reader, there is no boundary to the universe, and the universe is seemingly infinite.

However, from outside of the universe, the balloon barrier is quite observable, as is the creation and destruction of the balloon. Religiously speaking, one plays the part of God when inflating a balloon, with complete control over what goes through that one small – and controllable – opening of the balloon.

It is this opening through which dependency injection of resources can occur. When an application needs access to a file system, for example, it supplicates the future creator of its universe by enumerating its requirement as part of its application definition. These requirements are collected by the compiler and stored in the resulting application binary; any attempt to create a container for the application will require a file system resource to be provided.

And there are two ways in which such a resource can be obtained. First of all, the resource is defined by its interface, so any implementation of that interface, such as a *mock* file system or a fully emulated – yet completely fake! – file system would do. The second way that the resource can be obtained is for the code that is creating the container to have asked for it in the same manner – to declare a dependency on that resource, and in doing so, force its own unknown creator to provide the filing system as an answer to prayer.

As the saying goes, it’s turtles all the way down. In this case, the outermost container to be created is the root of the container hierarchy, which means that if it requires a filing system, then the language runtime must inject something that provides the interface of a filing system, and that resource that is injected might even be a representation of the actual filing system available to the operating system process that is hosting the language runtime.

And here we have a seemingly obvious contradiction: What is the difference between a language that attempts to protect resources by hiding them at the bottom of a sandbox container, versus a language that provides access to those same resources by injecting them into a container? There are several differences, but let’s start with an obvious truth: Perfection in the design of security is difficult to achieve, and even harder to prove the correctness of, so it is important to understand that this design does not itself guarantee security. Rather, this design seeks to guarantee that only one opening in the balloon – and anything that is injected through that opening – needs to be protected, and the reason is self-evident: Transitive closure. By having nothing naturally occurring in the language runtime that represents an external resource, there is simply no surface area within the language runtime – other than the injected dependencies themselves – that is attackable.

Second, the separation of interface and implementation in the XVM means that the implementation of the resource is not visible within the container into which it is injected. While this pre-introduces a number of language and runtime concepts, the container implementation only makes visible the surface area of the resource injection *interface* – not of the implementation! This holds true even with introspection, and furthermore the injected resources are required to be either fully immutable, or completely independent services.

Third, this design precludes the possibility of native code within an Ecstasy application; native functionality can only exist outside of the outermost container and thus outside of the language runtime itself, and can only be exposed within the language runtime via a resource injected into a container, subject to all of the constraints already discussed.

Lastly, as has been described already, the functionality that is injected is completely within the control of the injector, allowing the requested functionality to be constrained in any arbitrary manner that the injector deems appropriate.

While it is possible to introduce security bugs via injection, the purpose of this design is to minimize the scope of potential security bugs to the design of the relatively small number of interfaces that will be supported for resource injection, and to the various injectable implementations of those interfaces.

## On Processor Performance

There exists no shortage of opinions on the topic of what aspects are the most important in a system of execution. One voice will claim that only performance matters, while another will suggest that it no longer matters at all. One voice will claim that achieving efficiencies in development is far more valuable, while another will insist that predictability and stability in execution is critical. Opinions morph with time, as the reality of the physical units of execution evolves and the conceptual units of design are ever the more realized in languages and libraries.

Nonetheless the state of the art today bears the hallmark of a path followed far beyond its logical conclusion. In 1977, John Backus raised an early warning in his ACM Turing Award lecture:

Surely there must be a less primitive way of making big changes in the store than by pushing vast numbers of words back and forth through the von Neumann bottleneck. Not only is this tube a literal bottleneck for the data traffic of a problem, but, more importantly, it is an intellectual bottleneck that has kept us tied to word-at-a-time thinking instead of encouraging us to think in terms of the larger conceptual units of the task at hand. Thus programming is basically planning and detailing the enormous traffic of words through the von Neumann bottleneck, and much of that traffic concerns not significant data itself, but where to find it.

While programming advances have largely digested and expelled the explicit concerns of store-addressing and word-at-a-time thinking, these advances have been repetitively accreted onto a burial mound whose foundation remains a von Neumann architecture. Perhaps the success of that underlying architecture is the result of natural selection, or perhaps we have only inertia to blame. In any case, the evolution of concurrent multi-processing and distributed systems has stretched the von Neumann architecture past its effective limits. Specifically, it appears that the recent growth in the extent of the now automatically-managed store has occurred at a pace well beyond the increase in performance of the heart of the von Neumann machine: the processor. Whether this imbalance can be rectified by further technological accretion or by the adoption of a fundamentally new execution architecture is yet to be seen, but regardless: The inevitable and predictable increase in performance that has become the opiate of an industry has taken a sabbatical, and may have accepted an early retirement altogether.

There has existed a loose historic alignment in the growth of processor performance, memory capacity, memory throughput, durable storage capacity, durable storage throughput and network throughput. This relatively consistent growth has allowed a general model of assumptions to be perpetuated throughout hardware architectures, operating systems, programming languages and the various resulting systems of execution. Now we find that model to be threatened by the failed assumption that processor performance will increase at a rapid and relatively predictable rate.

To maintain the façade of progress, explicit hardware parallelism has emerged as the dominant trend in increasing processor throughput. Symmetric Multi-Processing (SMP) has a relatively long history in multi-CPU systems, but adoption of those systems was hobbled both by high prices and a lack of general software support. In the 1990s, the advent of the World Wide Web propelled multi-CPU systems into the mainstream for servers, but it is the recent, seemingly instantaneous and near-universal commoditization of multi-core CPUs that has finalized the dramatic shift from a focus on processor performance to a focus on processor parallelism. Further compounding the adoption of multi-CPU and multi-core systems are various technologies for Concurrent Multi-Threading (CMT), which enables a single CPU core to execute multiple threads concurrently. In aggregate, since the turn of the millennium, parallelism has increased from one to dozens of concurrently executing threads in an entry-level server, while the performance of an individual processing unit has only increased by only a few times. Looking forward, processor performance is now expected to improve only incrementally, while the level of parallelism is continuing on an exponential curve.

Since overall processing throughput has continued to increase at a dramatic pace not dissimilar from its historic trend, this shift from performance to parallelism could be safely ignored but for one problem: The von Neumann architecture is bound to a individual processing unit, and thus has nearly halted its forward progress in terms of the throughput of a single thread of execution. This means that for the first time in software history, existing programs do not run significantly faster on newer generations of hardware unless they were built to explicitly take advantage of thread parallelism, which is to say unless they were built assuming their execution would occur on multiple von Neumann machines in parallel. Since the art of programming is expressed almost entirely in imperative terms, and since the imperative nature of programming languages is based on the von Neumann architecture, we have managed to accumulate generations of programs and programmers that are hard-wired to a model that has at least temporarily halted its forward progress.

As a result, computing devices are providing increases in processing throughput that can only be consumed by parallelism. It is obvious that this mandates support for parallelism in any new system of execution, but there is a far less obvious implication of critical importance. Parallelism increases throughput only to the extent that coordination is not required among the threads of execution, and coordination is required only for mutable resources that have the potential to be shared across multiple threads of execution. In common modern systems of execution such as Java and C#, explicit parallelism is provided by threads of execution, each representing the state of a single von Neumann machine, but those machines collectively share a single store. Compounding the coordination overhead for the store is the prevalence of automatic management of the store, referred to as Garbage Collection (GC), which unavoidably requires some level of coordination. While GC algorithms have advanced dramatically in terms of parallelism, the remaining non-parallelized (~~and possibly non-parallelizable~~[[7]](#footnote-7)) portion of GC is executed as if by a single thread, which is to say that the runtime behavior of these systems will periodically halt in order to garbage-collect the shared store. The unavoidable conclusion is that growth in the shared store without the corresponding increase in processor performance has lead to unavoidable and growing pauses in the execution of the parallelized von Neumann machines.

A series of advances in GC algorithms have thus far masked this inevitable consequence, but the advances are already showing diminishing returns, while the upward pressure on the size of the store has not abated and the dramatic progress of processor performance has not resumed. Ultimately, a single shared mutable store must be avoided, and the design of a runtime system must reflect that requirement. The XVM design is intended to fully address this challenge, and does so by decomposing the problem space along a number of naturally occurring fault lines. First, the XVM organizes scopes of linking, loading, and execution – referred to as *containers* – in a hierarchical manner; second, within that hierarchy, scope of execution is further localized into individual, conceptually-independent von Neumann machines – referred to as *services* – within which all allocations occur; and third, that only immutable state can *escape* the execution scope of a service.

These decisions allow dynamic memory allocation to be managed within (scoped to) a particular service, and the resulting garbage collection to be performed entirely within the context of individual services, without requiring the coordination of other services or containers. The two exceptions to this are escaped immutable data, and the reclamation of services and containers themselves. In the case of escaped immutable data, and precisely because the data (the objects) are immutable, the memory can be garbage collected without any coordinated halt of execution[[8]](#footnote-8). Furthermore, the escaped immutable data can be organized within the container hierarchy at the level to which the data has escaped, or alternatively can be managed in a single global immutable store.

In the case of containers and services, each mark phase of each actively executing service also marks any services that it in turn can reach, again without any coordinated halt of execution; unreachable services are then collected asynchronously.

The concept of localizing allocations in order to localize GC is not new. Systems built around an explicit threading model have employed *escape analysis* in order to determine which allocations can safely be performed using a thread local allocator (such as a slab allocator), and which allocations need to be made from a shared store. This represents a hierarchy with a fixed depth of two: Global and thread local. While a dramatic improvement over a single shared store, it still implies a global stoppage for GC execution of the shared store.

The primary benefit to GC of the localization of the store is that a significant portion of overall GC execution can be localized entirely within each thread of execution, and further localized within the bounds of a particular service. Additionally, having stores that are localized to each service enables the system to exactly measure and meter – in real time – the amount of memory that is consumed by each service, and in aggregate by each container. Lastly, a range of optimizations are available to the data managed in a thread-localized store: the memory containing the data can be allocated, accessed, manipulated, and freed efficiently without any hardware-level concurrency control, and generated native code can be optimized specifically for cache locality.

Second, by organizing memory hierarchically in a manner corresponding the runtime container model, a service or an entire container can be discarded in its entirety, because no objects outside of that hierarchical scope can have a reference into the memory managed within that scope. In other words, the cost of “yanking” an entire sub-portion of the hierarchical runtime system is nominal, and the results are deterministic.

An additional benefit is that machine code optimized for single-threaded execution has dramatic performance advantages compared to machine code that is concurrency-safe, even when concurrency control is optimized using the latest hardware capabilities such as compare-and-swap (CAS) and other “lockless” instructions. Mutable data structures that are localized to a single thread of execution allow the resulting execution to approach the theoretical maximum efficiency of the underlying hardware.

The concept of GC optimizations based on immutability is also not new. Several GC implementations have leveraged memory protection barriers to protect memory regions being compacted as if the data were immutable, allowing application execution to proceed (in the absence of a fault) with the GC operating concurrently. Significantly, explicitly immutable data can be compacted without protection, because both the old (pre-compaction) and new (post-compaction) copies of the data are valid – being identical! – thus enabling the application to continue to execute concurrently and correctly while referring to either copy arbitrarily, deferring the housekeeping task of updating any pointers that point to the lingering old copy, and deferring the recycling of the memory region that was compacted. As an added benefit, the GC of regions of data known to be immutable can be performed by any thread of execution, or even a separate thread dedicated to GC.

## On Immutability

In an object-oriented system, immutability refers to the prohibition to alter the state of an object. It turns out that many data types are naturally immutable; consider the number 42 for example[[9]](#footnote-9) – it is always the number 42! Other data types are naturally mutable; consider the concept of a *variable* for example – its very purpose is to be able to vary! Many data types are naturally immutable, and even with mutable data types, it is often desirable to be able to make specific instances immutable on demand.

The XVM explicitly supports immutability. Immutability has several benefits, notably: Predictability, thread/concurrency safety, security, and available optimizations. Predictability is one of the greatest benefits of good design, and immutability supports predictability by providing data types that are truly constant – *at runtime!* For example, when an object exposes its state, it often does so by exposing immutable data types so that its internal state cannot be directly altered; without explicit support for immutability, one of two things could occur: Either the mutable state of the object would be exposed, breaking encapsulation, or a copy (or other representation) of the mutable state would need to be created on demand to safely expose that state, which is expensive in terms of both space and time, not to mention complexity. Immutability provides a simple way to ensure that the state of an object *cannot* change, addressing each of these concerns.

With respect to thread and concurrency safety, an immutable object provides the same state regardless of the concurrency of access to the object, because changes to the state of the object are explicitly prohibited. It is precisely because of this explicit contract that the only state that can be visible to more than one thread of execution in the XVM is immutable state; immutable objects can be safely used without concurrency control and without relying on memory barriers.

Using immutability for security is powerful, but it is important to understand that security as a topic is simply another facet of predictability, and as a result, the same concepts and conclusions apply. Specifically, when immutability is an intrinsic axiom of a system, and thus cannot be circumvented, it becomes a powerful tool for reasoning about how aspects of the system will operate, and those assumptions become trusted building blocks for building secure software.

Regarding optimizations, the explicit knowledge of immutability conveys a number of significant advantages to a system of execution. As described previously, for example, explicitly immutable data enables a number of potential optimizations for the purpose of garbage collection, and immutability obviates the need for the types of concurrency control used with mutable objects. Immutability also supports both pass-by-reference and pass-by-value semantics *transparently*, because the underlying value is itself immutable, and thus the substitution of any duplicate copy of the value has no behavioral consequence. (Put another way: In the XVM, one cannot determine whether an object is being passed by reference or by value *if that object is immutable*.)

In addition to object immutability supported by the language runtime, there are two related XVM capabilities worth enumerating. The first is support for lazily initialized state as part of an otherwise-immutable data structure, specifically, by means of a function with presumed-idempotent behavior. Such a capability allows the evaluation of time-expensive computations and/or the allocation of space-expensive data structures to be deferred until actually requested. One common example is the hash function calculation for complex data types, which is often assumed to be expensive enough to defer, but whose immutable value, once computed, should be saved for subsequent use.

The second capability is language-level support for designing a non-mutating reference to an underlying data structure that may itself be mutable. In other words, it is desirable to be able to support multiple references to the same object, such that a reference explicitly omits mutating operations and the exposure of mutable state. There are many examples of mutable data that must be generally protected from mutation, but which the owner of the data may need to be able to mutate; in this case, the reference through which mutations are prohibited is called a *read-only* reference. This capability does not rely on immutability, but rather relies on the careful design of a programmer, and as a result does not provide the types of guarantees that actual immutability can confer. However, with careful use, it is a useful tool for selectively hiding mutability.

## On References

The family of languages influenced by C++ share an implicit trait: A compile-time knowledge of accessibility. For example, it is the C++ compiler that prevents access to private members, and the compiler that allows access for a friend. Subsequent languages, like Java, built on this model, and carry sufficient information in the compiled form of the language to allow the accessibility implied by the compiled code to be validated for a second time at runtime, to prevent an end-run around the language’s security features. One by-product of this design is the ability to use a single pointer to reference an object, and more specifically, a pointer that is – in C++ parlance – a Vtable\*\* (a pointer to a pointer to a virtual function table).

In the simplest terms[[10]](#footnote-10), one can imagine an object as the combination of a structure (such as a C struct) and functions that operate on that structure. For each particular type (known as a class), the Vtable\*\* approach arranges those function pointers in an array, and orders those function pointers in the array to be consistent with other compatible types, thus supporting the polymorphic behavior of compatible types (via the common sub-sections of those arrays). A pointer to the array of functions is stored in the first field of the struct, which means that a pointer to the struct will also act as a pointer to the functions associated with that struct (hence the object pointer being a Vtable\*\*). Invocation of any of those functions requires that a pointer to the struct (which pointer is named this) be passed to the function, thereby allowing the function to access and modify fields in the struct, and allowing it to invoke other functions against that same struct, all without statically binding to either the address of the struct or to the addresses of the other functions related to the struct.

From a mechanical-simplicity and efficiency standpoint, the benefits of this model are difficult to overstate; it is powerful, and it is elegant. However, there are several specific costs to account for as well. First, the type system defines accessibility in a static manner, predicated on the class of the referrer and the class of the referent, and how those two classes relate. Consider C++ protected members, which are accessible to sub-classes, or Java package-private members, which are accessible to any class within the same package (a hierarchical name-space). Second, the facets that a class exposes are statically fixed, such as the set of members declared as public, allowing any referrer that obtains a reference to exploit runtime capabilities such as Java’s *reflection* to fully examine public members, access any public data fields, and invoke any public methods. While allowing arbitrary access to public members does not seem at first to be a concern, it can be; consider that members of an interface are always public in an implementing class, which leads to the third issue: It is not possible for an object to selectively expose its members, for example providing different interfaces to different referrers.

This lack of selective exposure can lead to complexity, particularly when the goals of rich functionality and security collide. One need look no further than Java’s serialization capabilities, in which an object must be able to expose its state to a serialization mechanism, but that same state – often private in nature – must be protected from all other referrers. The solution to conflicting goals is to create exceptions to the rules. Unfortunately, each exception to a rule contributes to the complexity of the system, and faults (vulnerabilities) inevitably emerge from complexity.

Ecstasy employs a different model entirely, by *conceptually* composing each reference from two pieces of information: An object’s identity, and the type (or interface) exposed through the reference. This approach is a synthesis of several concepts:

* A *type* is defined not by its name nor by its implementation, but solely by its member capabilities, exposed as a set of methods and properties;
* A *reference* encodes both the capabilities that are selectively permitted, and the object identity against which those capabilities may be invoked;
* While each class of objects has a number of predefined sets of capabilities, such as public, protected, and private interfaces, a set can also be arbitrarily composed;
* The permission to access the capabilities of an object is part of (as if *encoded in* and *defined by*) the reference information itself, and is neither the result of static compile-time checks nor dynamic security checks;
* It is always permissible to reduce (to *mask*) the type of a reference by removing properties and/or methods, but the opposite (to *reveal*) is forbidden, except *at or above* the point in the container hierarchy where the object’s class was loaded (similar to the *ring model* in a CPU).

There are several effects of this design decision:

* Any number of different references can exist for the same underlying object, each of which can have a different set of exposed methods and properties;
* There are cases in which it will be possible for the XVM to provide a reference to a purely derivative object of an existing object *without actually allocating a derivative object*, by instead creating a references whose type is the type of the derivative object, but whose implementation is modified to rely on the original object. For example, the same *identity* can be used to represent a particular object *and* an array of objects containing only that particular object, by encoding that same identity with two different *types* into two different references[[11]](#footnote-11).

Lastly, the ability to *mask* and *reveal* portions of functionality in a reference is a fundamental security mechanism. When a reference from outside of a container is injected into that container, that reference is *masked* as the injection type (such that only the methods declared on the interface being injected will be accessible), and within that container, it is then impossible to *reveal* additional methods on that reference. However, within a container, for classes defined within that container, it is always possible to reveal additional methods on references to objects created within that container. In other words, modules loaded within a container are not protected from other modules loaded *within the same container*, and objects instantiated within a container are not protected from other objects instantiated *within the same container*. What does this mean in practice? Among other things, it means that within the container that defines a class and instantiates an object of that class, terms like private and protected are useful only for hiding complexity, not for securing the contents of an object.

## On Portability

Similar to the goals of the Java Virtual Machine specification, portability is a primary consideration for the XVM. Specifically, the XVM design targets the efficient implementation on three major instruction sets (x86/x64, ARM A32/A64, and wasm[[12]](#footnote-12)) and five major operating systems (Android, iOS, Linux, macOS, Windows); it is assumed that support for any other modern environment is practical as a result.

|  |  |  |
| --- | --- | --- |
| Instruction Set | Operating System | Example |
| ARM | Android | Phone, tablet, desktop, kiosk, embedded device |
| ARM | iOS | iPhone, iPad, Apple TV |
| x86/x64 | Linux macOS Windows | Desktop  Server |
| wasm | \* | Browser |

The XVM specification defines a portable binary (aka *byte code*), and the behavioral requirements to execute such a binary. The XVM is not constructed around nor constrained by a particular word size, pointer size, endian byte ordering, or signed/unsigned support of the host CPU; all encodings in the portable binary are either octet (8-bit byte) or variable-length integers.

The portable binary was designed as an input for native compilation, such as a Just In Time (JIT) compiler. Additionally, care was taken in the design of the XVM to ensure that it allowed for Ahead Of Time (AOT) compilation. One explicit non-requirement of the portable binary is efficient interpretation; while an XVM interpreter is possible (one was implemented as a proof-of-concept during development of the XVM specification), the design of the portable format does not lend itself to efficient interpretation.

There is one root[[13]](#footnote-13) module, the Ecstasy.xtclang.org module, which provides the fundamental building blocks of the type system; all XVM types are derived from the contents of this root module. Furthermore, a module cannot contain native (external) code, so the root module contains a complete implementation of the types contained therein. This implies a completely self-referencing type system composition, which while a correct assumption, is also an impossible conclusion[[14]](#footnote-14).

While the root module itself is identical (portable) across hardware and software platforms, the implementation of the XVM varies – potentially dramatically – across the same. The XVM is responsible for replacing select portions of the self-referential implementation in the root module with native implementations that support the specified contracts; however, *which* portions get replaced will vary depending on a combination of the goals of a particular XVM implementation and the capabilities of the hardware and software combination that the XVM is targeting. Furthermore, *which* portions get replaced is (and *should be*) unknown by a programmer working in Ecstasy.

On one hand, it is highly desirable to be able to provide a truly minimal set of native implementations in the execution system itself, and to be able to implement all other data types and operations by composing from that minimal set. This approach permits a minimalistic implementation of the XVM. Similarly, it is highly desirable to be able to provide a richer set of native implementations as part of the execution system itself, in order to leverage additional capabilities provided by the runtime environment, the operating system, and the underlying hardware. In other words, the ability to trade off complexity, size, and performance is based on the ability to define standardized capabilities as abstract data types that can be implemented either as an native part of the execution system or in software that is executed by the execution system. While this design aspect is visible in as few places as possible, it deeply permeates the design of the XVM specification and the root module.

To accomplish this goal, the root module provides a rich set of abstract types, defined as part of the XVM specification, and a reference implementation for all aspects of those types is provided. It is *possible* to create one XVM implementation that relies *almost* entirely on the reference implementation of these types (thus requiring only a nominal set of operations to be implemented in native code), and another XVM implementation that directly implements a dramatically larger portion of the root module in native code (hopefully achieving higher performance with less memory usage).

With the advent of the CLI specification, best known in the Microsoft .NET CLR incarnation, and to a lesser extent in the open source Mono project, a new facet was added to the concept of virtual machine portability: Explicit multiple-language support. There are inevitable trade-offs in creating an execution system that is intended to support multiple languages; generally, an execution system is optimized for a specific language, and the implementation of any other language trades off between execution performance and the adherence to the “foreign” language specification.

One can conceivably implement any programming language on any Turing-complete execution system, resorting to full runtime interpretation if necessary; the obvious trade-off is correctness versus space/time efficiency. With respect to multiple language support, the primary goal of the XVM is to enable the efficient implementation of a *specific class of language* that can be described by its attributes: modular, object-oriented, composable, single-dispatch, immutable message passing, concurrent, thread-safe, with automatic memory management. Alternatively, the class of language could be described as an incremental evolution of the Java and C# languages, with support for type composition, a reified generic type system with transitive closure, a single (object) type system, and a recursively consistent (hierarchical) runtime model, while explicitly eschewing support for concurrent access to mutable data. While it is desirable that other classes of languages be efficiently implementable on the XVM, it is an explicit non-requirement that the semantics of those languages be intrinsically supported by the XVM where those semantics differ from those of the targeted specific class of language. Or, more succinctly: The XVM is designed to run the Ecstasy language, and its design allows for relatively efficient execution of languages with similar runtime models, but its design contains nothing that explicitly supports languages other than Ecstasy.

## On Startup Time

One of the lesser design goals for the XVM is to be able to achieve near instantaneous startup time. As simple as this goal sounds, the complexity of the type system works against it.

Consider a side effect of the core JVM type system being largely implemented on top of the JVM: The large number of classes required for the simplest “hello world” application. Just as in the children’s song[[15]](#footnote-15) that iterates over the body a bit at a time, like “*the knee bone’s connected to the shin bone*” and so on, the JVM is forced to load a relatively large graph of its type system as a result of the dependencies among its most basic types. Loading a type likely requires I/O, parsing, validation, internal type definition, the execution of any specific type initialization code, and the recursive loading of types on which the loaded type depends. While loading a relatively large graph of a type system is acceptable for a long-running process (which can amortize that loading cost over hours, days or months), it is far more intrusive a cost for a short-running process.

The XVM design further exacerbates this condition by guaranteeing – with transitive closure – the validation of an entire type system before it is activated within a runtime container. That means that the entire type system must be loaded and evaluated before the first line of code can be executed.

To address this challenge, several aspects of the XVM design are relevant:

* Within a given container, the type system of that container is immutable; loading the type system within a container thus always results in an identical type system;
* The elimination of global data structures (and in particular, mutable data structures) dramatically simplifies coupling across the type system, simplifying initialization;
* Dependency injection largely eliminates lazily-initialized language sub-systems, such as I/O;
* AOT compilation allows a container’s type system, or even an entire application, to be compiled to executable code and related structures in advance; and
* Explicit version support in the portable binary allows the runtime to detect changes that violate any of the above assumptions.

The design is intended to allow an XVM implementation to instantiate a new container with a specified set of modules, with a cost that is *on the same order of magnitude* as loading and initializing a shared library in a running C/C++ application.

## On Openness

Undoubtedly, the largest change in software in the past two decades has been cultural and not technological. The term “open” had long been used for marketing commercial software that had some slight yet often only theoretical potential for interoperability with other commercial software. Today, most core software components, libraries, operating systems – and even many applications – are available in complete source code form for use under an *open source* or *software libré* license, and many of the specifications and standards – including languages and execution systems – that enable interoperability are similarly open and available. From an economic standpoint, it would appear that the demand for a fundamental set of software standards and components being available as a public good now out-weighs the cost of creating and managing that public good (even in some cases lacking any consistent centralized authority!), and the cost of reverting to private goods for that fundamental set of software standards and components is unacceptable for all but the most especial of requirements.

It is in this spirit that the XVM specification is made available, with its ideas and concepts inspired from others’ open work, and – if any prove worthwhile – its ideas and concepts freely available for re-use and recycling as the reader sees fit.

## Credits

None of these ideas occurred in a vacuum. The initial inspiration for this effort was the Java Virtual Machine, which was groundbreaking in its timely fusion of brilliance and pragmatism, and which is described in beautiful detail by the Java Virtual Machine Specification and the related Java Language Specification. Older languages such as C, C++, Python, and Erlang left their own indentation on our thinking as well. Many concepts of the XVM will be familiar to Smalltalk programmers, who to this day righteously espouse that all subsequent programming advances were already present in and perfected by Smalltalk, which unfortunately met its untimely demise as the result of a vast right-wing conspiracy. More recently, languages such as C#, Ruby, Scala, Go, Ceylon, and Kotlin have stretched our imaginations, each with their own pragmatic preview of a potential programming future, and with systems of execution such as the Sun Hotspot JVM, Microsoft CLR, Google V8, and LLVM managing to occasionally exceed even the scope of imagination with their practical ingenuity.

# Background

Ecstasy and the XVM exist in a technology field filled with existing knowledge an experience. It was our intention, in designing Ecstasy, to leverage as many existing concepts and design idioms as we could from the most-used languages and technologies, so that someone with existing knowledge in field would feel immediately at home both reading and writing in the Ecstasy language.

This section attempts to correlate technical terms and concepts used in the Ecstasy langue and the XVM with similar terms and concepts that are already broadly used in the field.

## A Brief Introduction to Object-Oriented Programming

A number of languages are described as Object-Oriented, which means that the language is intended to allow the structure of a program to be neatly divided into discrete units that each define a specific portion of the program’s data and the logic that is closely associated with that data; these units are referred to as *classes*. To explain some terms courtesy of an overused analogy, consider a dog named Spot and a cat named Fluffy: Spot is an object whose class is dog, while Fluffy is an object whose class is cat. The term *class* represents the definition of the thing, while the term *object* represents one of those things; sometimes the term *instance* is used, as in: Spot is *an instance of* dog.

Now that the obligatory animal analogy has been made, let us instead consider a graphics system on a device or computer, starting with the concept of color. One representation of color is called RGB, in which three separate numeric values for red, green, and blue are mathematically combined into a single numeric value that represents the color; perhaps the simplest implementation of RGB values represents the three color channels as 8-bit integers, and the combined RGB value as a 24-bit integer. In an object-oriented language, one can define a class named Color whose properties are Red, Green, Blue, and RGB. The notion of *encapsulation* allows the detail of how that information is stored inside a Color object to be hidden; for example, a Color could have three separate integers that it stores its red, green, and blue information in, or it could have one integer that it stores the combined RGB value in, or it could use some other representation of the data altogether. Regardless, from any point of view outside of the Color class, the *state* of a color object is easily represented by the *properties* of the object in a manner that is consistent with the description of what a color is, such that values of the Red, Green, Blue, and RBG properties are consistent.

Another valuable aspect of object-oriented programming is *polymorphism*, which is when different classes of objects share some common programming *interface*, and thus are substitutable to some degree. Extending our previous example, a second class representing color could be created in a manner that would be substitutable in some situations for the first, by providing the same programming interface as the Color class provides, but potentially with a different data structure and implementation. For example, some JPEG and video formats use the YCbCr representation of color that is composed of a luma value (Y), and chroma values for both blue difference (CB) and red difference (CR). Since a YCbCr value can be converted to and from an RGB value, a second Color class (perhaps named YCbCrColor) could be defined that provides properties of Luma, BlueDiff, and RedDiff, but also provides properties of Red, Green, Blue, and RGB – just like the original Color class. Again, as with the previous example, encapsulation allows the class to hide the specific manner in which the object manages its internal information (also called its *state*) behind its programming interface, while exposing that information as both an YCbCr value *and* as an RGB value. Polymorphism allows an instance of the new YCbCrColor class to be used where the program was expecting an instance of the original Color class.

One way in which the YCbCrColor class could have been defined is through the use of *inheritance*, which is when a new class is defined by starting with the definition of an existing class, and then adding things to it. Inheritance is quite powerful, particularly because it allows the re-use of existing class definitions, and because it provides a natural means of supporting polymorphism. In our example, if YCbCrColor is inherited from Color, then it would automatically inherit the entire programming interface of Color, and an instance of YCbCrColor could be transparently used in place of an instance of Color, because YCbCrColor *is a* Color. However, as powerful as inheritance may seem, it has often been perceived to be overused and abused as a tool; as a result, developers have started to “favor object composition over class inheritance”[[16]](#footnote-16).

While inheritance provides an “*is a*” relationship, composition provides a “*has a*” relationship. For example, instead of the YCbCrColor class inheriting from the Color class, it could simply have a property of the Color class; in other words, the programming interface for YCbCrColor could include a way to obtain a separate object that *is a* Color value, just like it includes a way to obtain Luma, BlueDiff, and RedDiff values.

A third option is to use interface inheritance, which allows the programming interface from an existing class to be re-used while not inheriting other aspects of the existing class definition, such as the manner in which data is structured and managed. For example, a programming interface called Color could be defined that has a single property RGB, whose value is in the form used natively by the device, such as a 24- or 32-bit integer or floating point value. Any number of class definitions could then *implement* the Color interface, thus allowing instances of those various classes to be utilized any time a Color is required. For example, the Color class from our original example could be renamed as RGBColor, and both it and the YCbCrColor class could be made to implement the new Color interface. Similarly, the interface could be implemented by additional class definitions, such as a YUVColor class (the color value used in PAL televisions), and a YIQColor class (the color value used in NTSC televisions).

Conceptually speaking, a programming interface is composed of the accessible *properties* (state) of a class and the *methods* (behavior) that the class exposes; collectively, these are the *members* of the interface. Some programming languages allow any class to act as a substitute for another class, as long as it has the same members as the class that it is substituting for; this is an example of *dynamic typing*, and is often specifically referred to as *duck typing* because “if it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.” Other languages require that a class explicitly declare the names of the programming interfaces for which it can act as a substitute; generally speaking, this is an example of *static typing*. However, it is important not to approach terms such as static and dynamic typing too rigidly, because these concepts represent a continuum of possibilities, as opposed to some single binary choice. Furthermore, every programming language has many different aspects, each of which can be located at a completely different point on that continuum.

In object oriented programming, each class is a specific data type; in our examples, each of the various color-related classes and programming interfaces is its own separate data type. When a programmer wishes to create a class (e.g. a Array) that will manage, or contain, or operate upon some unspecified second class (e.g. the class of things that will go into the Array), that second class can be represented by a place-holder, which is called a *type parameter* (e.g. an ElementType). A class that has one or more type parameters is called a *parameterized type*, or simply a *generic type*. We have already discussed “*is a*” and “*has a*” relationships; a generic type supports an “*of a*” relationship. Consider the array, which is perhaps the most basic of data structures, supported even by assembly languages. An array is a contiguous sequence of a specific number of elements of a specific type, each of which can be loaded from or stored into the array using the element’s index, which is an absolute position in the array.

While static typing makes it possible to specify that one must pass a single Color to a particular method, consider a method that accepts multiple Color objects to be passed to it in an array. Without generic types, the programmer would have to define the method to accept *any* Array (whose contents are of an unspecified type), or alternatively could create a specific type – perhaps ArrayOfColors – that can hold only colors. A generic type is an alternative to creating a separate specific type (such as ArrayOfColors) for each possible “*of a*” relationship; instead, a single type (Array) is defined with a place-holder, such as Array<ElementType> which uses ElementType as a place-holder for some unspecified type. Using a generic type would allow the method in our example to be defined to accept a parameter of type Array<Color>, which is read as “an Array of Color”. The benefit of generic types is the extension of type safety (and any other facilities that rely on or can benefit from type knowledge) to any classes that support one or more “*of a*” relationships. This is also referred to as *parametric polymorphism*, which refers to both the *parameterized type* aspect, and polymorphism, which is a direct benefit of using parameterized types (generic types) as the basis for generic programming.

One of the valuable capabilities provided by inheritance, and perhaps one of the causes for the overuse of inheritance, is the ability to define state and behavior in one class that is then inherited by many different classes. Elimination of redundant code (*a la* “cut and paste”) is a noble goal indeed, and is at the core of the “Don’t Repeat Yourself” (DRY) principle[[17]](#footnote-17), which states: “Every piece of knowledge must have a single, unambiguous, authoritative representation within a system.” In addition to inheritance, you can see this principle at work in programming interface definitions, polymorphism, and generic types. The concept is well described by Steve Smith[[18]](#footnote-18):

Of all the principles of programming, Don't Repeat Yourself (DRY) is perhaps one of the most fundamental. The developer who learns to recognize duplication, and understands how to eliminate it through appropriate practice and proper abstraction, can produce much cleaner code than one who continuously infects the application with unnecessary repetition.

Every line of code that goes into an application must be maintained, and is a potential source of future bugs. Duplication needlessly bloats the code-base, resulting in more opportunities for bugs and adding accidental complexity to the system. The bloat that duplication adds to the system also makes it more difficult for developers working with the system to fully understand the entire system, or to be certain that changes made in one location do not also need to be made in other places that duplicate the logic they are working on.

Repetition in logic can take many forms. Copy-and-paste[..] is among the easiest to detect and correct. Many design patterns have the explicit goal of reducing or eliminating duplication in logic within an application. In fact, the formulation of design patterns themselves is an attempt to reduce the duplication of effort required to solve common problems and discuss such solutions.

A powerful mechanism for achieving DRY in object-oriented programming is the *mix-in* (or *mixin*), which in many ways is like a class definition, but one that can be “mixed into” (i.e. “added to”, “applied to”, or “incorporated into”) other class definitions. Using the concepts of generic programming, a mixin *applies to* some yet-to-be-specified second class that provides a specified programming interface, allowing the logic of the mixin to be constructed around the abilities of that programming interface. Then, any time that mixin’s logic is desired in another class, the mixin can be applied to that class; while various other mechanisms provide “*is a*”, “*has a*”, and “*of a*” relationships, a mix-in satisfies the “*needs a*” relationship. Continuing our earlier color example, some graphics systems support the notion of color transparency, which is sometimes expressed in the range of 0% to 100%, but is often encoded as an 8-bit value prepended to a 24-bit RGB value, creating a 32-bit integer value. A ColorTransparency mix-in could be defined that applies to any Color, adding a Transparency property that holds a value representing the transparency level, and overriding the RGB property to incorporate the transparency information into the resulting value.

This “brief” introduction was not intended to be an exhaustive explanation of object oriented programming; hopefully, it has introduced a number of terms and concepts, and with enough context to provide a basis for understanding some of the constructs in the Ecstasy language.

In the end, these concepts are simply mechanisms for organizing functionality together into logical units, in manners designed explicitly to minimize redundancy and errors, and to maximize readability and understandability. Regardless of the language, the computer underneath still executes in the same manner, and its concepts are far simpler, and its mechanisms far fewer in number. Each discrete piece of logic is, at some level, an assembly language function, and accessing that logic is, at some level, nothing more than a call instruction. It is that world that each of these object oriented constructs must translate into, such that every conceivable operation can be expressed as nothing more than a function call, and the magic, if any, is in doing so efficiently without the developer ever being aware of that original intent.

## Types

A system of execution has at its core a concept of data types. The XVM relies on a core set of types that are – by self-reference – necessary to bootstrap the XVM and those types themselves; these types, being intrinsic to the XVM, are referred to as *intrinsic types*, and are defined in the root module, Ecstasy.xtclang.org. As described previously, the real implementation of these types may differ from implementation to implementation of the XVM specification, but there is an ironclad guarantee: That these types are always available.

The XVM uses a pure object type system, and only an object type system. As the words would imply, a pure object type system does not include support for types that are *not* object types; traditionally such non-object types are known as *primitive types*. The lack of a primitive type system is familiar to Smalltalk programmers, and has emerged as a general trend in newer object-oriented languages. There are several reasons for the emergence of pure object type systems, but the fundamental reason is that it dramatically simplifies a language by reducing the number of intrinsic type systems – typically from two (objects and primitives) to just one (objects). The result is uniformity and simplicity, and thus elegance.

However, there are legitimate technical reasons why primitive types still exist in most contemporary languages. First of all, they are compact: While an object often consumes 16 bytes of memory at a minimum, a primitive byte is still one byte. Second, an object is likely to incur an additional memory dereference on every access, because objects tend to be accessed by pointer; any virtual behavior incurs at least one more dereference on top of that. Third, it is relatively easy to generate efficient native code for a high-level language that uses primitive types, as if the code had been written in a primitive language such as C. Finally, there are native hardware accelerators, such as Single-Instruction Multiple-Data (SIMD) co-processors[[19]](#footnote-19) and CPU instruction sets that require values to be arranged in a **very** explicit layout composed solely of explicitly primitive (native) types.

Even though raw performance concerns have diminished as the processing throughput of CPUs has increased, it is still crucial for the success of an execution system to carefully consider performance trade-offs. Nonetheless, there are two fundamental realizations that enable the abandonment of a primitive type system without significant concern for a negative performance impact. First, the use of runtime profiling information to optimize code has opened new possibilities by illustrating how conceivable-yet-daunting performance optimizations can be safely realized; this trend is successfully and dramatically illustrated by the Sun (now Oracle) Hotspot JVM. Second, in a pure object-oriented type system, primitive types are translated to objects that are immutable and whose identity corresponds only to the represented value; this knowledge unleashes a slew of potential optimizations. For example, multiple copies of the object can be safely created without concern of identity confusion, since any two objects corresponding to an identical primitive value are in fact the same *singleton* object whether or not their location in memory is the same; as a result, it is conceivable to arbitrarily optimize to using a native representation on the stack, in a register, or as an immediate value, thus achieving the same performance benefits in the executing native code as would have been possible with an explicit primitive type system.

The loss of support for a primitive type system without the loss of potential performance is not accidental. In order to achieve this, there are several common programming concepts that must be hidden[[20]](#footnote-20) such as memory location and method of allocation, and several programming concepts that must be made explicit such as immutability and equality (sameness). These concepts are not limited to those types formerly known as primitive; the associated benefits can be realized with any intrinsic or user types that share these same attributes.

## Objects

In a traditional imperative language like C, complex data types are typically represented as *structs*; a struct is considered a complex data type because it is composed of any number of *fields*, each of which has its own data type. It is possible both to instantiate a struct “on the stack” without dynamic allocation of memory, and to instantiate a struct “on the heap” using dynamic allocation of memory. Generally, *pointers* to those structs are used as an efficient mechanism to pass a reference-to-a-struct to a function, to store a reference-to-a-struct in a variable or any data structure, etc., without actually having to copy or move the struct itself. Following a pointer to get to the pointed-to struct is called de-referencing the pointer.

In an object-oriented language like C++ (which itself extends the C language), a *class* is a data type that extends the concept of a struct, but by default hides the fields of the struct as *private* members. Like C, it is possible to pass a C++ object using a pointer to the object – which would require de-referencing the pointer to access the object – but it is also possible to pass a C++ *reference* to the object, thus partially hiding the concept of a pointer. Furthermore, code that is part of the definition of the class is automatically invoked when an instance of the class is constructed and destructed. All of these aspects of C++ are designed to support and encourage encapsulation, which is a key tenet of object-oriented languages. C++ also enables the definition of the class itself to be parameterized using *templates*; these templates are used to extend otherwise-generic class definitions with additional compile-time type safety and type-specific optimizations.

Java removed a number of the C and C++ language capabilities altogether:

* Structs;
* Pointers;
* Multi-dimension arrays;
* Control over the memory layout of classes;
* Stack-allocation of objects;
* Creating multiple objects with a single allocation;
* Explicit object destruction and deallocation; and
* Derivative patterns such as RAII.

The result was an elegant compromise that allowed developers to focus on a pure object type system, with a few obvious exceptions, such as:

* The non-object type system represented by primitive values;
* The not-entirely-pure object type system represented by array types, including some “split personality” types like “arrays of primitive values”;
* Global functions (functions not related to object instances) retained via the static keyword;
* Non-virtualized object instantiation (since an exact type must be specified following the new keyword).

Java does not expose pointers to the programmer; instead, Java has *object references*. These act very similarly to pointers, but with a few notable differences:

* The reference is an opaque data structure, in that its only value is the referent itself;
* The memory address of the referent object is not exposed;
* The reference itself cannot be manipulated or modified (e.g. pointer arithmetic);
* A reference is strongly typed and (mostly) type safe;
* There is exactly one reference value that does not have a referent, which is the null value, which has well defined behavior.

On the other hand, like C and C++, Java retained the ability to compare two references for reference equality, but then added a few additional related and (in retrospect) unwise contracts, including:

* Java references are associated with a native *hash-code value* that originally was thought to be “free”, since the value was the memory address (or a direct transformation thereof); maintaining this contract ended up adding an extra four bytes to every object (which was eventually optimized out of many objects at a high cost of optimization complexity).
* Java references served as the basis for Java’s native (i) mutual exclusion, (ii) thread parking, and (iii) thread unparking mechanisms, referred to as (i) synchronized methods and blocks, (ii) wait, and (iii) notify; again, this had the unintended consequences of making every object a potential mutex and parking queue, adding some amount of storage overhead to the header of every object, and adding dramatic optimization complexity to the JVM to reduce the overhead of those unintended consequences.

(Java is truly an engineering marvel, but the greatest feats of the Java team are found in the ways that they have managed to overcome the seemingly insurmountable challenges of their own making. While Google’s V8 engine has accomplished much the same for the morass of JavaScript, the V8 team can at least disclaim the various design flaws that they are so ingenuously[[21]](#footnote-21) compensating[[22]](#footnote-22) for.)

The XVM uses objects and references, in much the same manner as the JVM and the CLR before it.

## Terms

**Object** – Every value, every input, every output – quite literally every *thing* in the XVM is an object. Objects are *managed* by the XVM: They are *instantiated*, they exist for a period of time, and they are *garbage collected* when they are no longer used. An object is a conceptually discrete unit of state and behavior.

**Class** – Every object is “*of a*” class, and that class defines the state that can be held by an object, and the behavior of an object.

**Reference** – Every reference is “*to an*” object (the *referent*). Every interaction with an object occurs through a reference to that object, including interactions that an object has with itself. A reference has a *type*, and that type defines the set of interactions that are possible with the referent object through the reference.

**Type** – A type defines a set of interactions that are possible with an object, expressed in the form of properties and methods.

*Every reference has a type. Every object has a class.*

this – Every object has a reference to itself; the reference to self is the object’s this reference[[23]](#footnote-23).

**Meta** – Every object has an associated representation of its design-time and run-time meta-data; this representation is provided by the object’s *meta* object.

**Struct** – Every object has an associated representation of its structure that contains its state; this representation is the object’s *struct* object. For each property of the object[[24]](#footnote-24), there exists a *field* in the object’s struct that contains the reference held by the property.

**Template** – The unit of design as provided by a developer. Templates are combined to form a class, using information encoded in the template, and following rules defined in this specification. A template is “of a” template category; there are eight template categories, corresponding to the keywords module, package, class, const, enum, service, mixin, and interface. A template has an identity, and defines properties and methods. A template can also define constants, functions, and can contain child templates.

**Property** – A property is a discrete unit of object state; an object’s state is accessed and modified via properties. A property has a constraint type and an identity, and holds a reference to the property’s value. The property value is stored in the object’s structure in a *field*, and is TODO virtual on the object

**Method** – A method is a named piece of functionality – a *behavior* – that can be applied to an object; that object is called a *target object*. As a result of combining multiple templates in order to form a class, a method may represent a *sequence* of method implementations, because each template can contribute an implementation for a given method. (This concept is called a *virtual method*, and the sequence of method implementations is called a *virtual call chain*.) A method, when *bound* to a target object, produces a *function*.

super – For each method implementation in a virtual call chain, the next implementation in the chain is called the *super method* (which implies that the last method implementation in the chain does not have a super method). In each method implementation that has a super method, a function reference named super is available to invoke the super method.

**Constant** – A constant is a named immutable value. The constant may be a *compile-time* constant, if its value is determinable by the compiler, or it may be a *run-time* constant, if its value is computed at run-time, which occurs no later than the first time that the value is used.

**Function** – A function is an *implementation* that can be *invoked*; it represents a sequence of *instructions*. A named function can be defined as part of a template, in a manner similar to a method; unlike a method, a function is not virtual, and a function does not get bound to a target object. A function takes some number of *parameters*, and yields some number of *return values*. When one or more function parameters are *bound* to argument values, a new function is produced (with that many fewer parameters). A function, when *invoked*, executes the instructions in the implementation, and yields the return values.

## Meta Model

While the intrinsic type system is intended to provide the fundamental building blocks from which the software developer can construct their own arbitrarily elaborate types, it also provides the software developer with a definition of the XVM itself, as it is running, including the type system. In other words, the type system includes intrinsic types that define and describe the type system. The use of a self-referential and self-defining type system is not new, but Ecstasy and the XVM carry the concept to its (recursively) logical conclusion.

Based on the concepts already discussed, it should come as no surprise that a class in the XVM is itself an object, and the same applies to all of the constituent pieces of a class, such as methods, properties, and so on. Even a function is an object, which is an incredibly simple and powerful way of expressing a potential action (the function) as a “thing” that can be stored off, passed around, compared with other functions, and so on – all in a transitively closed, type-safe manner. Obviously, a type itself is an object, which itself has a type, which itself is an object, and so on to infinity.

There are a few aspects of the XVM that are treated as if they were closely guarded secrets – things that in theory could easily be exposed through a meta-model, but are not. Things like how an object is laid out in memory, and how that memory is managed – whether it is a dynamic allocation, or on an execution stack, and so on. As described previously, these are conscious choices, because by hiding this information, the XVM is free to organize (and even dynamically reorganize) the native structures and code that constitute a running system.

One of these closely guarded secrets is the structure of an object reference, and that is because it is desirable to avoid native structures and pointers altogether for many common types, such as a Boolean or an Int. By hiding what a reference *is*, the runtime can pretend that there is an object when in reality there is none, and can pretend that there is a reference to that object, when in reality there may not be. In other words, the XVM can choose to implement an object as an object, but it is also free to implement an object in any way that it finds more desirable, such as implementing an object as a machine-native value[[25]](#footnote-25).

Amazingly, that object (that does not actually exist), referenced via a reference (that does not actually exist), can have code that is running with a this (of that non-existent reference), being passed around from place to place, having its virtual methods executed – and all of the language and runtime contracts still hold!

And while the details of the object’s structural organization are purposefully hidden, and while the references themselves are not directly manipulable, the XVM meta-model carefully exposes the concepts of the *referrer* (the origination point of a reference) and the *referent* (the target object of a reference), and magic ensues. Every value, every instantiated object, every method or function result, every property, every calculation – every *object*! – only exists because it is referenced; once it is no longer referenced, it ceases to exist. We often describe that *state of being referenced* as “someone holding onto an object”, and in reality that simply means that a reference to that object is stored in at least one reachable place in the system, be that a variable in a method, a property on another object, an element in an array, and so on. That “place where the reference is stored” is called a Ref.

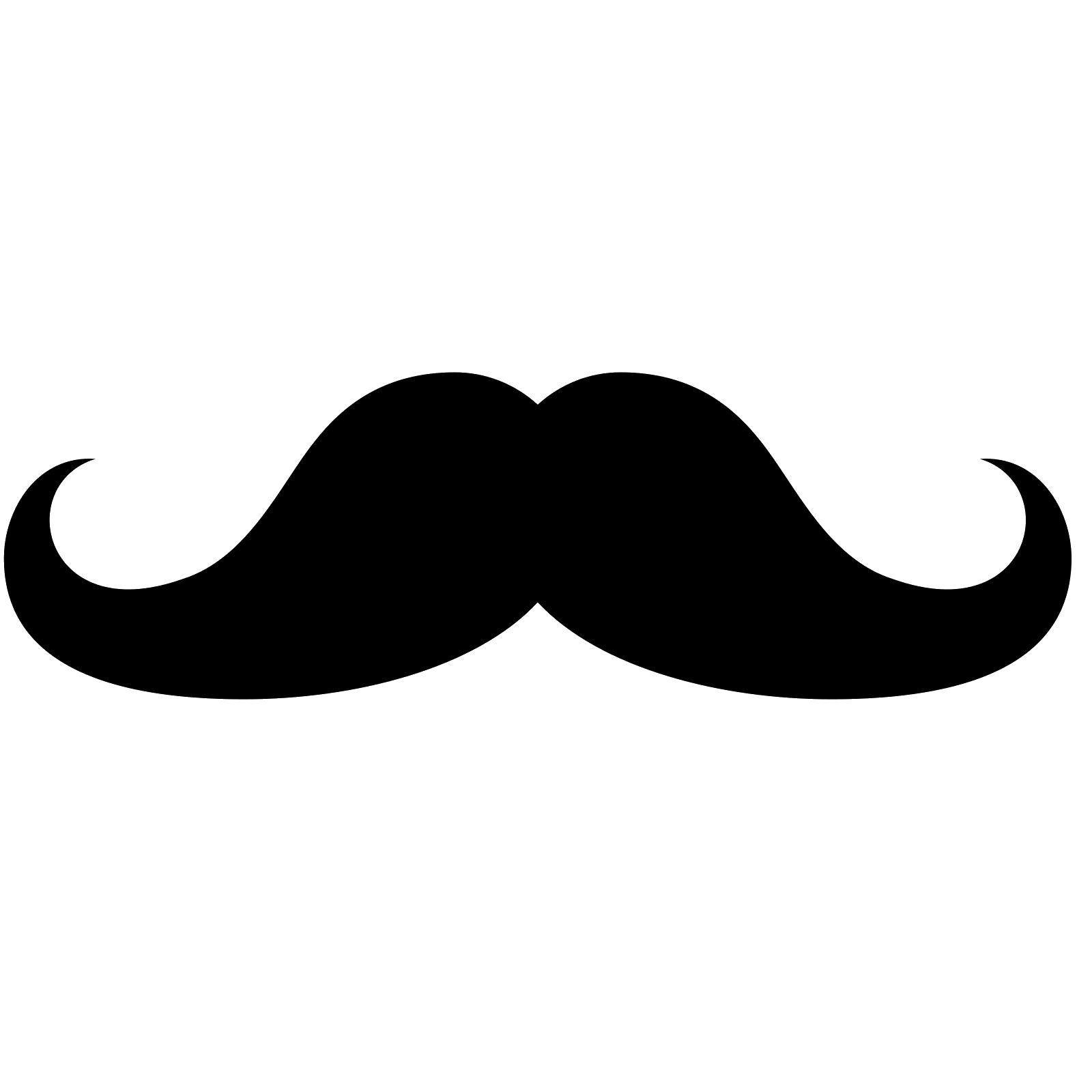
The magic is that each Ref is itself an active, extensible, programmable part of the XVM, and not just some passive address in memory. Every place that an object can be held, and every place that an object can be accessed, there is a Ref, and to operate on a reference, one must ask the Ref to get it, and to change the reference, one must ask the Ref to set it. When a local variable is declared, the declaration is specifying the class of the Ref that will hold the variable’s value. When a property is declared, the declaration is specifying the class of the Ref that will hold the property’s value – and the property declaration may even be declaring a new class of Ref that exists just to manage that one property’s value!

As already described with respect to objects themselves, the contracts for Ref are carefully designed, so that the runtime can pretend that there is a Ref when in reality there may be none, and in the case that a developer ever asks for a reference to the Ref itself, it will always be there.

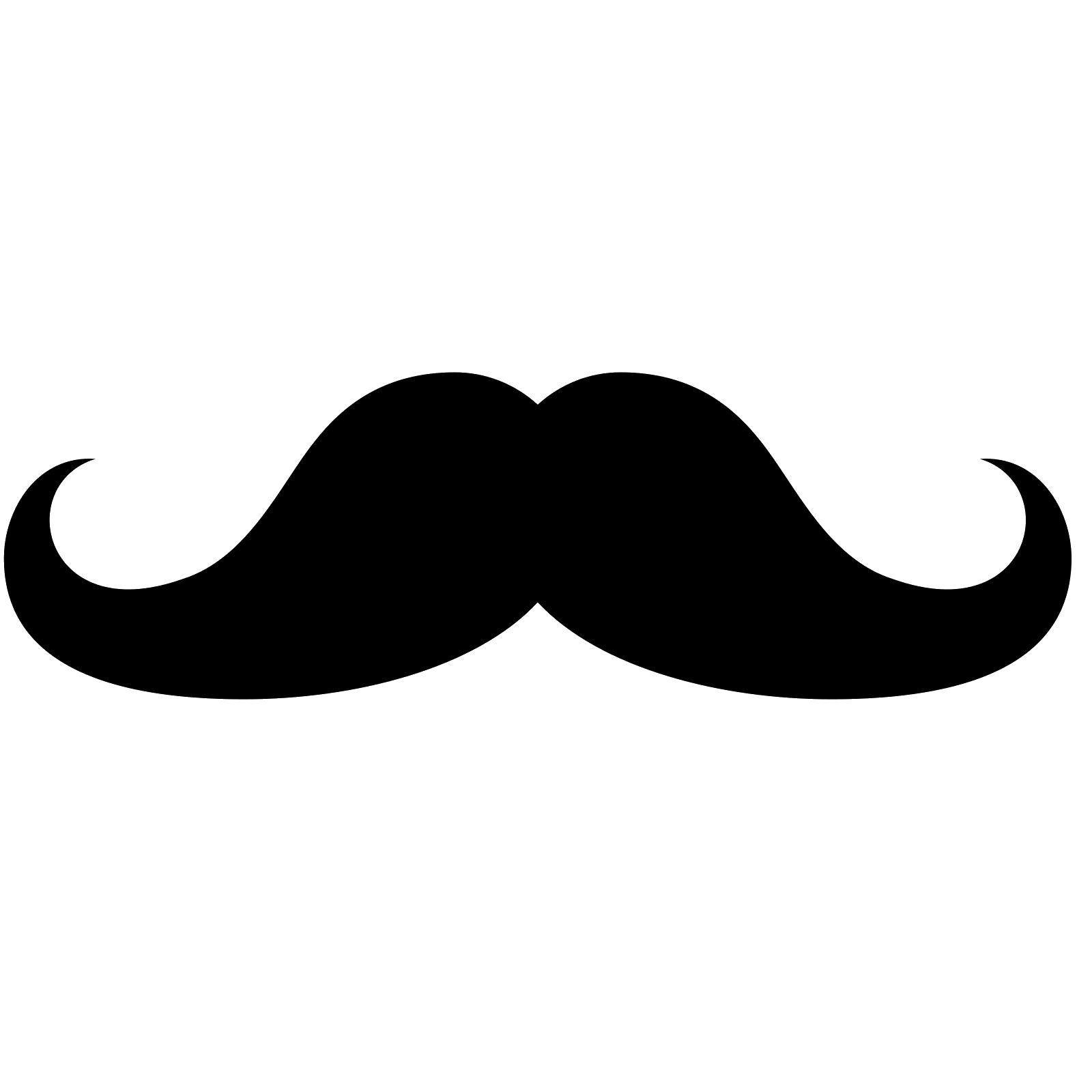
## Mixins and Annotations

The discussion of Ref is an ideal segue to the topic of mixins, and their use in annotations. To understand the concept of a mixin, it helps to start by considering the traditional concept of class inheritance. Inheritance starts with something that we don’t wish to alter (or ruin!), yet it’s something that we wish to use as the basis for something new and different, so instead of altering the original, we place a “pane a glass” over it, and do our work there:

**+**



**=**



By composing in this manner, we still have the original, and we also have a new *derivative* work (the “Monad Lysander”) that exists – without destroying or defiling the original! If these were classes, we would describe the original as the *base class*, and the result as the *derived class*, with the relationship being that the derived class *inherits from* the base class.

Implementation inheritance has long been regarded as an incredibly powerful tool, but it is also regarded as being one of the most fragile forms of composition in an object-oriented system. In the case of Monad Lysander, all of that careful movembrish work is tightly coupled to the intricate underlying details of Mona Lisa. Should we wish to enhance any other masterpiece in a similar manner, inheritance would force us to reproduce all of that painstaking follicular work, almost certainly relying on the “cut & paste” design pattern.

Remembering back to the goals of Ecstasy, we posited: “*a language should enable a developer to locate each piece of design and logic in its one best and natural place*.” That “pane of glass” represents that one best and natural place, but only if it can be applied to any face in any picture, and that ability is what defines a *mixin*.

*A mixin is a template that can be incorporate into any other template that satisfies the mixin’s type constraint, using that type as its joinery into the resulting class.*

In our example, the *type constraint* is “has an upper lip without a mustache”, and that is the point to which the mixin is affixed. Since the mixin’s type constraint is known, the compile-time type for its this is the type specified in the type constraint.

Unlike inheritance, in which a template specifies the template from which it must derive, a mixin can apply to any template that *is of* the necessary type. This allows the mixin – that one “*piece of design and logic*” – to be used over and over again, anywhere that it is needed. Furthermore, any number of mixins can be incorporated into a resulting class, allowing common functionality to be split out into discrete units (each a mixin), and then aggregated together as that functionality is needed.

But perhaps most importantly, a mixin can be incorporated into an existing class *by annotation* to form a new class on the spot; consider this line of Ecstasy code:

return new **@Persistent** Employee(id, name);

And at this point, we must digress, because it’s simply too obvious that one could benefit by adding capabilities – such as persistence – in a business application to business classes – such as an “Employee”. While novel, indubitably clever, and intriguingly interesting, such a solution is also eminently predictable. What makes mixins amazing is that a mixin can be incorporated into a class that is not even known.

Consider a variable, for example. We can prove that a variable exists, precisely because it holds a value in a predictable manner. We can even obtain the Ref object for that variable, and since we have stated as an axiom that an object has a class, we know therefore that the variable itself – being an object! – has a class. But the developer does not “new” that class to make a variable; rather, the existence of that variable just happens – exactly when, and where, and *as* it is needed.

And here, incorporated into that unknown class, is where the mixin performs its magic:

**@Future** HttpResult result = svc.handle(request);

This declaration is *not* incorporating a Future mixin into an HttpResult; rather, it declares that *the class of the variable itself* incorporates the Future mixin. Indeed, the result variable[[26]](#footnote-26) (augmented with its Future functionality) is now capable of being treated as an asynchronous future:

**&**result.handle(e -> genErrorPage(request, e))

.passTo(result -> send(request, result));

The same annotation capability is also used to declare properties. A number of Ref-altering annotations (e.g. @Future, @Atomic, @Lazy, @Soft, and @Weak) are included in the root module, but if the need arises, it is startlingly simple to write a new Ref-altering annotation of your own.

Lastly, annotations can apply to either the variable itself, or to the type contained in the variable; in this example, the variable has a @Lazy mixin, and it holds a reference to an @Unchecked Int object:

@Lazy(() -> 2+2) @Unchecked Int x;

# Ecstasy: The XTC Language

The XTC[[27]](#footnote-27) language is the reference language implementation for the XVM.

## Lexical Structure

XTC code is at its most fundamental level a sequence of Unicode code-points, referred to as *characters*. When a sequence of these characters is stored in a file, it is typically stored in an *encoded* format, for example using the UTF-8 encoding. A number of possible encodings exist for sequences of Unicode characters, but the details of those encodings are outside of the scope of this specification. All encodings must support at least the Unicode code-points in the range U+0000 through U+007F, corresponding to the ASCII character set, and encodings may support up to the entire range of legal Unicode code-points, which at the time of writing is U+0000 through U+10FFFF.

The XTC source code being lexically analyzed is called the *Input*. For purposes of lexical analysis, the *Input* is considered to be a character stream, and that stream is read from left to right.

The first stage of lexical analysis tracks the *line number* and the *line offset* in order to identify exact locations within the source code; both the line number and line offset are zero-based. Any *location* within the source can be identified by a combination of the line number and the line offset, while a selection of the source is identified by a starting location (inclusive) and an ending location (exclusive). The location tracking uses the definition of the *LineTerminator* sequence to determine when to increment the line number and reset the line offset:

*LineTerminator:*

U+000A

U+000B

U+000C

U+000D U+000Aopt

U+0085

U+2028

U+2029

Each *RawCharacter* in the character stream is evaluated to determine if it is the beginning of a *LineTerminator* sequence; if it is, then reading past the *LineTerminator* sequence causes the line number to be incremented and the line offset to be reset to zero. Reading any other *RawCharacter* (that is not the beginning of a *LineTerminator* sequence) causes the line offset to be incremented and does not change the line number. Note that in the case of the two-character *LineTerminator* sequence also known as “CR/LF”, the line number and line offset are undefined in between the contiguous characters U+000D and U+000A.

The second stage of lexical analysis translates Unicode character escape sequences into characters. A Unicode character escape sequence can occur at any point within the source, and must specify a legal Unicode code-point.

*UnicodeCharacterEscapeSequence:*

\u *hex-digit* *hex-digit* *hex-digit* *hex-digit*

\U *hex-digit* *hex-digit* *hex-digit* *hex-digit* *hex-digit* *hex-digit* *hex-digit* *hex-digit*

Here are a few examples of how specific characters could be encoded using a Unicode character escape sequence:

|  |  |  |
| --- | --- | --- |
| Character | Code Point | Escape Sequence |
| $ | U+0024 | \u0024 or \U00000024 |
| ¢ | U+00A2 | \u00A2 or \U000000A2 |
| € | U+20AC | \u20AC or \U000020AC |
| 𤭢 | U+24B62 | \U00024B62 |

Since the stream is being analyzed as if it is being read left to right, one character at a time, the character that results from a Unicode character escape sequence is not further evaluated to determine if it could be part of another Unicode character escape sequence; for example, the Unicode code-point U+005C is the character ‘\’, but the source “\u005cu005c” translates simply to the six characters ‘\’, ‘u’, ‘0’, ‘0’, ‘5’, ‘c’, and not further to the single character ‘\’. Similarly, since the Unicode character escape sequences are evaluated after the line number and line number processing, the Unicode character escape sequence “\u000a” does not affect the line number; instead, since the escape sequence is six characters long, it increases the line offset by six, even though a subsequent stage of lexical analysis would only encounter a single character. Note that neither a recursively encoded format nor an escapable format is supported; “\uu0024” is translated simply as the seven characters ‘\’, ‘u’, ‘u’, ‘0’, ‘0’, ‘2’, ‘4’, and “\\u0024” is translated as the two characters ‘\’, ‘$’.

The third stage of lexical analysis converts the residual *Input* stream of characters into a stream of *InputElements*, which are *WhiteSpace*, *Comments,* and *Tokens*, prefaced with an optional *BeginningOfFile* serving as Unicode byte-ordering indicator, and terminated with an optional *EndOfFile*; the *WhiteSpace* and *EndOfFile* elements are discarded.

*Input:*

*BeginningOfFileopt InputElementsopt EndOfFileopt*

*BeginningOfFile:*

U+FEFF

*InputElements:*

*InputElement*

*InputElements InputElement*

*InputElement:*

*WhiteSpace*

*Comment*

*Token*

*EndOfFile:*

U+001A

Whitespace is defined as the union of the traditional ASCII whitespace characters and the three Unicode “separator” categories. Here is the entire table of whitespace characters:

|  |  |  |  |
| --- | --- | --- | --- |
| Code Point | Decimal | Name | Description |
| U+0009 | 9 | HT | Horizontal Tab |
| U+000A | 10 | LF | Line Feed |
| U+000B | 11 | VT | Vertical Tab |
| U+000C | 12 | FF | Form Feed |
| U+000D | 13 | CR | Carriage Return |
| U+001C | 28 | FS | File Separator |
| U+001D | 29 | GS | Group Separator |
| U+001E | 30 | RS | Record Separator |
| U+001F | 31 | US | Unit Separator |
| U+001A | 26 | SUB | End-of-File, or “control-Z” |
| U+0020 | 32 | SP | Space |
| U+0085 | 133 | NEL | Next Line |
| U+00A0 | 160 | &nbsp; | Non-breaking space |
| U+1680 | 5760 |  | Ogham Space Mark |
| U+2000 | 8192 |  | En Quad |
| U+2001 | 8193 |  | Em Quad |
| U+2002 | 8194 |  | En Space |
| U+2003 | 8195 |  | Em Space |
| U+2004 | 8196 |  | Three-Per-Em Space |
| U+2005 | 8197 |  | Four-Per-Em Space |
| U+2006 | 8198 |  | Six-Per-Em Space |
| U+2007 | 8199 |  | Figure Space |
| U+2008 | 8200 |  | Punctuation Space |
| U+2009 | 8201 |  | Thin Space |
| U+200A | 8202 |  | Hair Space |
| U+2028 | 8232 | LS | Line Separator |
| U+2029 | 8233 | PS | Paragraph Separator |
| U+202F | 8239 |  | Narrow No-Break Space |
| U+205F | 8287 |  | Medium Mathematical Space |
| U+3000 | 12288 |  | Ideographic Space |

Since the longest possible translation is used at each stage of lexical analysis, *WhiteSpace* and *Comments* are often necessary for separating two tokens that could otherwise be translated as a single, larger token.

*WhiteSpace:*

*WhiteSpaceElement*

*WhiteSpace WhiteSpaceElement*

*WhiteSpaceElement:*

*SpacingElement*

*LineTerminator*

*SpacingElement:*

U+0009

U+001C

U+001D

U+001E

U+001F

U+0020

U+00A0

U+1680

U+2000

U+2001

U+2002

U+2003

U+2004

U+2005

U+2006

U+2007

U+2008

U+2009

U+200A

U+202F

U+205F

U+3000

*LineTerminator:*

U+000A

U+000B

U+000C

U+000D U+000Aopt

U+0085

U+2028

U+2029

*Comment:*

*EndOfLineComment*

*EnclosedComment*

*EndOfLineComment:*

// *InputCharactersopt*

*InputCharacters:*

*InputCharacter*

*InputCharacters InputCharacter*

*InputCharacter:*

*RawCharacter* except *LineTerminator*

*EnclosedComment:*

/\* *EnclosedCommentTail*

*EnclosedCommentTail:*

\* *AsteriskCommentTail*

*NotAsteriskCharacter EnclosedCommentTail*

*AsteriskCommentTail:*

/

\* *AsteriskCommentTail*

*NotAsteriskOrSlashCharacter* *EnclosedCommentTail*

*NotAsteriskCharacter:*

*RawCharacter* except \*

*NotAsteriskOrSlashCharacter:*

*RawCharacter* except / or \*

*RawCharacters:*

*RawCharacter*

*RawCharacters RawCharacter*

*RawCharacter:*

any legal Unicode code-point except *BeginningOfFile* or *EndOfFile*

It should be obvious that the “control-Z” character used for the *EndOfFile* production, which is an historical anachronism, is only going to be permitted to occur at the end of the file. Similarly, the Unicode byte-ordering indicator character used for the *BeginningOfFile* production is only permitted to occur at the beginning of the file. Note that the *BeginningOfFile* production can be utilized in a number of different Unicode encodings, including UTF-8, UTF-16, and UTF-32 encoded files, and in either little-endian or big-endian byte order for both the UTF-16 and UTF-32 formats, and thus represents a number of different byte sequences that one may encounter at the beginning of a file. Any program operating on XTC source files should recognize and support no less than these five Unicode encodings and the plain ASCII encoding.

Since the stream is being analyzed as if it is being read left to right, and since the *EndOfLineComment* is seeking only for a *LineTerminator* or the end of the stream, any “//”, “/\*”, or “\*/” character sequences encountered while producing an *EndOfLineComment* will be ignored. Likewise, since the *EnclosedComment* is seeking only for the “\*/” character sequence, any “//” or “/\*” character sequences encountered while producing an *EnclosedComment* will be ignored. By logical inference, comments cannot be nested.

## Lexical Literals

There are several fundamental literal forms that are handled by the lexical stage of compilation, including numeric literals and character string literals.

For numeric literals, Ecstasy supports base-2 (bit/binary), base-8 (octal), base-10 (digit/decimal), and base-16 (hexit/hexadecimal) values. Ecstasy requires a special prefix (the digit ‘0’ followed by a non-digit character) to indicate a value uses a base (radix) other than base-10[[28]](#footnote-28). Fractional values can also be specified in base-2, base-8, base-10, and base-16. Floating point values can have an exponent expressed in base-2, base-8, base-10, or base-16, and a floating point value that has both a mantissa and an exponent can use a different base for each.

*NumericLiteral:*

*IntegerLiteral*

*FloatingPointLiteral*

*IntegerLiteral:*

*Signopt UnsignedIntegerLiteral*

*Sign:* one of

+ -

*UnsignedIntegerLiteral:*

*BitLiteral*

*OctLiteral*

*DigitLiteral*

*HexitLiteral*

*BitLiteral:*

0 *BinaryIndicator Bits*

*BinaryIndicator:* one of

B b

*Bits:*

*Bit BitsOrUnderscoresopt*

*Bit:* one of

0 1

*BitsOrUnderscores:*

*BitOrUnderscore*

*BitOrUnderscores BitOrUnderscore*

*BitOrUnderscore:*

*Bit*

\_

*OctalLiteral:*

0o *Octals*

*Octals:*

*Octal OctalsOrUnderscoresopt*

*Octal:* one of

0 1 2 3 4 5 6 7

*OctalsOrUnderscores:*

*OctalOrUnderscore*

*OctalOrUnderscores OctalOrUnderscore*

*OctalOrUnderscore:*

*Octal*

\_

*DigitLiteral:*

*Digits*

*Digits:*

*Digit DigitsOrUnderscoresopt*

*Digit:* one of

0 1 2 3 4 5 6 7 8 9

*DigitsOrUnderscores:*

*DigitOrUnderscore*

*DigitOrUnderscores DigitOrUnderscore*

*DigitOrUnderscore:*

*Digit*

\_

*HexitLiteral:*

0 *HexIndicator Hexits*

*HexIndicator:* one of

X x

*Hexits:*

*Hexit HexitsOrUnderscoresopt*

*Hexit:* one of

0 1 2 3 4 5 6 7 8 9 A a B b C c D d E e F f

*HexitsOrUnderscores:*

*HexitOrUnderscore*

*HexitOrUnderscores HexitOrUnderscore*

*HexitOrUnderscore:*

*Hexit*

\_

*FloatingPointLiteral:*

*IntegerLiteral Exponent*

*Signopt FractionalLiteral Exponentopt*

*FractionalLiteral:*

0 *BinaryIndicator Bitsopt* . *Bits*

0 o *Octalsopt* . *Octals*

*Digitsopt* . *Digits*

0 *HexIndicator Hexitsopt* . *Hexits*

*Exponent:*

*ExponentIndicator IntegerLiteral*

*ExponentIndicator:*

*DecimalExponentIndicator*

*BinaryExponentIndicator*

*DecimalExponentIndicator:* one of

E e

*BinaryExponentIndicator:* one of

P p

There are three forms of character string literals in Ecstasy: The single character literal (in single quotes), the normal character string literal (in double quotes), and the free-form literal (inside of a two-dimensional textual frame). The first two are relatively self-explanatory for anyone with knowledge of contemporary languages, but the last is something altogether insane: Using the Unicode “box drawing” block of characters, and assuming a fixed-width font (for purposes of column alignment), a block can exist in Ecstasy code that contains free-form, multi-line text; for example:

String s = ╔═════════════════════╗  
 ║<html><body> ║  
 ║you could paste an ║  
 ║entire file in here ║  
 ║</body></html> ║  
 ╚═════════════════════╝;

The purpose of the free-form literal is to support the inclusion of entire text files, as-is, within source code files. This is particularly useful for including HTML, XML, JSON, and other templates as part of an application, located (embedded) right next to the logic with which they are used, without requiring elaborate concatenation in the source code itself. On the other hand, this capability does strongly indicate a need for a language-aware IDE that can manage the formatting of the embedded text.

*CharacterLiteral:*

' *SingleCharacter* '

*SingleCharacter:*

*InputCharacter* except \ or '

*CharacterEscape*

*StringLiteral:*

" *CharacterStringopt* "

*FreeformLiteral*

*CharacterString:*

*StringCharacter*

*CharacterString StringCharacter*

*StringCharacter:*

*InputCharacter* except \ or "

*CharacterEscape*

*CharacterEscape:*

\ \

\ "

\ '

\ b

\ f

\ n

\ r

\ t

*FreeformLiteral  
 FreeformTop FreeformLines FreeformBottom  
  
FreeformTop  
 Whitespaceopt FreeformUpperLeft NoWhitespace FreeformHorizontals NoWhitespace* **→**

*FreeformUpperRight Whitespaceopt LineTerminator*

*FreeformLines  
 FreeformLine  
 FreeformLines FreeformLine  
  
FreeformLine  
 Whitespaceopt FreeformVertical FreeformChars FreeformVertical Whitespaceopt LineTerminator  
  
FreeformChars  
 FreeformChar  
 FreeformChars FreeformChars  
  
FreeformChar  
 InputCharacter except FreeFormReserved or LineTerminator  
  
FreeformBottom  
 Whitespaceopt FreeformLowerLeft NoWhitespace FreeformHorizontals NoWhitespace* **→**

*FreeformLowerRight*

*FreeFormReserved  
 FreeformUpperLeft  
 FreeformUpperRight  
 FreeformLowerLeft  
 FreeformLowerRight  
 FreeformHorizontal  
 FreeformVertical  
  
FreeformUpperLeft  
 U+250C* ┌ *U+250D* ┍ *U+250E* ┎ *U+250F* ┏ *U+2552* ╒ *U+2553* ╓ *U+2554* ╔ *U+256D* ╭ *FreeformUpperRight  
 U+2510* ┐ *U+2511* ┑ *U+2512* ┒ *U+2513* ┓ *U+2555* ╕ *U+2556* ╖ *U+2557* ╗ *U+256E* ╮ *FreeformLowerLeft  
 U+2514* └ *U+2515* ┕ *U+2516* ┖ *U+2517* ┗ *U+2558* ╘ *U+2559* ╙ *U+255A* ╚ *U+2570* ╰ *FreeformLowerRight  
 U+2518* ┘ *U+2519* ┙ *U+251A* ┚ *U+251B* ┛ *U+255B* ╛ *U+255C* ╜ *U+255D* ╝ *U+256F* ╯ *FreeformHorizontals  
 FreeformHorizontal  
 FreeformHorizontals NoWhitespace FreeformHorizontal  
  
FreeformHorizontal  
 U+2500* ─ *U+2501* ━ *U+2504* ┄ *U+2505* ┅ *U+2508* ┈ *U+2509* ┉ *U+254C* ╌ *U+254D* ╍ *U+2550* ═ *FreeformVertical  
 U+2502* │ *U+2503* ┃ *U+2506* ┆ *U+2507* ┇ *U+250A* ┊ *U+250B* ┋ *U+254E* ╎ *U+254F* ╏ *U+2551* ║

## Lexical Symbols, Keywords, and Identifiers

Symbols are one or more characters (using the *“longest possible translation”* principle) that have a meaning in the language:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| : | { | << | ?: | &= | < |
| ; | } | >> | = | |= | <= |
| , | [ | >>> | += | ^= | > |
| . | ] | & | -= | &&= | >= |
| .. | + | | | \*= | ||= | ++ |
| ... | - | ^ | /= | := | -- |
| @ | \* | ~ | %= | ?:= | -> |
| ? | / | ! | <<= | == | \_ |
| ( | % | && | >>= | != |  |
| ) | /% | || | >>>= | <=> |  |

The following table contains all of the pre-defined keywords and identifiers (again, using the *“longest possible translation”* principle):

|  |  |  |
| --- | --- | --- |
| allow | function | static |
| as | if | struct |
| assert | immutable | super |
| assert:always | implements | switch |
| assert:once | import | this |
| assert:test | import:embedded | this:module |
| assert:debug | import:required | this:private |
| avoid | import:desired | this:protected |
| break | import:optional | this:public |
| case | incorporates | this:service |
| catch | instanceof | this:struct |
| class | interface | this:target |
| conditional | into | throw |
| const | is | TODO |
| construct | mixin | try |
| continue | module | typedef |
| default | new | using |
| delegates | package | val |
| do | prefer | var |
| else | private | void |
| enum | protected | while |
| extends | public | with |
| finally | return |  |
| for | service |  |

The last lexical element is the identifier. An identifier is composed of a sequence of characters[[29]](#footnote-29), the first of which must be a letter or an underscore, and the subsequent characters can include numbers and currency symbols:

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Unicode | Start | Trail |
| Letter | Lu Ll Lt  Lm Lo | ✔ | ✔ |
| Mark | Mn Mc Me |  | ✔ |
| Number | Nd Nl No |  | ✔ |
| Currency | Sc |  | ✔ |
| Underscore | U+005F | ✔ | ✔ |

The following table contains all of the pre-defined keywords and identifiers (again, using the *“longest possible translation”* principle):

*Identifier:*

IdentifierStart IdentifierFinish*opt*

*IdentifierStart:*

*Letter*

*\_*

*IdentifierFinish:*

*IdentifierTrail*

*IdentifierFinish IdentifierTrail*

*IdentifierTrail:*

*Letter*

*Mark*

*Number*

*Currency*

*\_*

*If you have built castles in the air, your work need not be lost; that is where they should be. Now put the foundations under them. – Thoreau*

1. XVM is an abbreviation for the XTC Virtual Machine [↑](#footnote-ref-1)
2. Such contracts are *software legacy*, meaning that the contracts cannot be altered without disclaiming the past; in other words, altering the contracts will break everything. [↑](#footnote-ref-2)
3. <https://www.joelonsoftware.com/2002/11/11/the-law-of-leaky-abstractions/> [↑](#footnote-ref-3)
4. <https://en.wikipedia.org/wiki/Sandbox_(computer_security)> [↑](#footnote-ref-4)
5. <https://en.wikipedia.org/wiki/Privilege_escalation> [↑](#footnote-ref-5)
6. The law of conservation of mass and energy, for example. [↑](#footnote-ref-6)
7. Since this was penned, Azul Systems (azul.com) commercialized a true, non-blocking software solution for large shared-mutable stores in their Java Virtual Machine implementation. [↑](#footnote-ref-7)
8. For example, an asynchronous copy-compacting collector allows multiple physical copies to represent the same object. Each individual service can perform its own mark phase, followed by a copy-compact phase (which can be performed by any thread, including a daemon), and any memory that is no longer used can be freed by the next mark phase. [↑](#footnote-ref-8)
9. See <http://hitchhikers.wikia.com/wiki/42>, [↑](#footnote-ref-9)
10. This example is intended to be illustrative, and should not be viewed as an authoritative explanation of object orientation unless you are attempting to annoy a Smalltalk programmer. [↑](#footnote-ref-10)
11. In this case, the term *type* is being used to indicate both the interface type and the implementation class. [↑](#footnote-ref-11)
12. WebAssembly is a portable machine code for web browsers: <http://webassembly.org/> [↑](#footnote-ref-12)
13. *“Root”* in the sense that it has no dependency on any other module. [↑](#footnote-ref-13)
14. Without expanding into a formal proof, the XVM itself provides no data types (such as primitive types) from which new types can be composed; thus, the root module defines types composed only from other types in the same module, which is by definition infinitely recursive. [↑](#footnote-ref-14)
15. This specification would lack provable transitive closure without a link to Dem Bones: <https://en.wikipedia.org/wiki/Dem_Bones> [↑](#footnote-ref-15)
16. From the book “Design Patterns: Elements of Reusable Object-Oriented Software”, GoF *et al* [↑](#footnote-ref-16)
17. From the book “The Pragmatic Programmer”, by Andy Hunt and Dave Thomas [↑](#footnote-ref-17)
18. With edits, licensed using [Creative Commons Attribution 3](mailto:https://creativecommons.org/licenses/by/3.0/us/) by Steve Smith [↑](#footnote-ref-18)
19. Most notably CUDA and OpenCL for driving massively parallel graphics processors (GPUs) [↑](#footnote-ref-19)
20. “*Hiding*” simply means that these aspects must not appear either explicitly *or implicitly* in a programming language contract. [↑](#footnote-ref-20)
21. <https://www.youtube.com/watch?v=UJPdhx5zTaw&t=48s> [↑](#footnote-ref-21)
22. <http://benediktmeurer.de/2017/12/13/an-introduction-to-speculative-optimization-in-v8/> [↑](#footnote-ref-22)
23. This is a simplification: There are actually five separate this references, each with a specific purpose. [↑](#footnote-ref-23)
24. This is a simplification: A property may or may not have a field in the object’s struct; a property that does not have a field is called a *calculated* property, which is to say that a property that is calculated may not require a field. [↑](#footnote-ref-24)
25. That same machine-native value can even be used to represent an array containing that object, or a tuple containing that object, or a function returning that object, all by varying the implied type in a reference, which itself may not even exist. The possibilities are mind-bending. [↑](#footnote-ref-25)
26. For a property or variable v of type T, the expression &v obtains a reference to the variable itself, which is of type Var<T> (or just Ref<T> if the property is read-only). The read/write interface Var extends the read-only interface Ref. [↑](#footnote-ref-26)
27. XTC is an abbreviation for XVM Translatable Code, and is also known as “Ecstasy” [↑](#footnote-ref-27)
28. Many languages allow an octal literal value to be specified by simply starting a value with ‘0’; this is one of the archaic constructs that Ecstasy does not support. A sequence of digits beginning with 0 is treated as base-10. [↑](#footnote-ref-28)
29. It is expected that these rules will be modified to allow the use of the standardized set of common emojis; see <https://unicode.org/emoji/charts/full-emoji-list.html> [↑](#footnote-ref-29)