CosmoS

Cosmos-B

IPC & SHARED MEMORY

*Cosmos-B is a managed code operation system designed to run on the Cosmos kit. This document describes the IPC and shared memory.*

IPC & SHARED MEMORY

Ben KloosterMan

*This document describes the IPC architecture and shared memory*

# 0.Overview

IPC in Cosmos B is probably the area where it will diverge the most from current systems. We aim for an order of magnitude improvement in IPC time due to the fact the only dispatch in the system is a Thread dispatch which we term “Task switch” to distinguish it from the current dispatch mechanisms even though the system is pre-emptive.

Key features of Cosmos-B IPC are

* No locking IPC , enhancing performance especially on mutli core systems
* IPC based APIs are easily extensible with good backward compatibility
* Asynchronous support again improving multi core performance
* Quick hand over to STPes waiting for messages reducing latency
* Bulk STPing of messages allowing for higher performance due to cache coherency and thread affinity.( At the cost of latency)
* Zero copy from Hardware to User for large messages or buffers.
* IPC and ABI ( kernel calls) are the same method
* More secure via pre secured point to point queues
* Less opportunities for tampering with data as messages are strongly typed and can be verified at a single point.

The extra performance will be used to improve multi Processor scalability, security as well as reliability. The IPC works especially well with functional languages which use immutable structures and are asynchronous.

Lastly the system will contain a full publish subscribe event system.

# 1.IPC/Shared Memory in MANAGED SYSTEMS

IPC in managed systems in theory is as simple as placing a reference to an object in a predetermined address and then the receiver can use that object. In practice there are a number of difficulties. Namely

* When the 2 parties have different GC’s the objects need to be carefully managed. This creates a range of issues.
  + - When one STP is killed what happens to the reference or object.
    - How do you handle references being added to and from different GCs into the object?
    - How do you handle GC compacting?
    - How do you track stack reference to the object.
    - How do we manage these references with memory checks
    - How to prevent dangling pointers eg duplicate a reference then set the original to null or tell the creator you’re finished.
* Thread Safety needs to be considered like current systems, which leads to locking.
* How to ensure security in that the object may lead to a significant security breach.

# IPC and SHARED MEMORY

Let’s look at the schemes

1. Copy on send requires small compact fixed length messages ( for efficiency) which interferes with a type safe system and general code quality. Best multi Processor support.
2. Parameter passing/RPC style results in chatty tightly bound interfaces that are hard to maintain. This especially a concern for an API which should evolve. In addition there are many points of entry . Not Asynchronous friendly and poor multi Processor performance. ( The stack is normally copied to the destination)
3. Sending a message provides type safety and easy maintenance ( a derived message can be sent in a new lib with no changes to the API) . Messages may be strongly validated at a single point of entry. The main issue is the message creation cost. Requires some form of shared memory.

The goal of the system is better security and reliability and the type safe messages clearly puts a strong case forward in this area. The use of shared memory is unlikely to have a significant cost (as it is established for the life of the SIA) . We hope that messages will become larger and more infrequent.

The cost of instantiating the message should be trivial as we use the NoConstructor Option , hence the cost is merely doing an Interlocked increment on the Nursery pointer and adding the type id to the object. For non Numa systems performance is likely to be the best for Asynchronous SIAs (like web servers) and especially multi cores.

# IPC -MESSAGE BASED, QUEUED and ASYNCHRONOUS.

One of the most important decisions to be made is whether to have Synchronous or Asynchronous IPC. And what method will be used for kernel calls (ABI).. As discussed the system will have no ABI (except for syscall) . All IPC will be point to point strongly typed messages and Asynchronous.

# API

The default API is simply Send(IKernelMessage msg) and Receive((IKernelMessage msg)

KernelMessage can be a struct , mutable or immutable depending on the implementation.

# IPC MethODS AND COMPARISON

There are 3 IPC methods These are

1. *Struct based IPC ( including events)*
2. *Sharing within the same GC*
3. *Sharing non mutable data [ Via ref counted Shared Heap]*

In all cases the actual method is just to send the reference the options above vary in the limits on the construction of the messages, where it is stored and how GC is dealt with.

# ISSUES WITH IPC

For non copy on write Shared Memory there are a number of issues namely.

All IPC needs to handle thread safety.

When using the Senders GC ( and the receiver has a different address space/GC)

* If it is managed by a GC you need to ensure it is not prematurely collected. Any such collection will break type safety and potentially security.
* You also need to consider what happens to the references when the sender(GC) is killed.
* You need to ensure the GC is not tied to other GCs which could propagate and cause a massive machine wise pause..
* The receiver may pass the reference to further STPes.
* The original sender eventually needs to be notified. ( hence this form of Shared Memory always requires a response)

When using a Shared heap.

* Who controls the heap
* State of data when crash occurs
* Duplication of references

Its worth restating the other forms of IPC/Shared Memory

* Copy on send , not really viable with buffers and other significant data
* Copy on write requires paging and VM or complex compiler changes and a bit of issues with managing the memory.
* Monolithic 1 GC and send reference approach is not viable due to massive GC pauses (which could result in HW IO issues)
* Traditional shared memory ( a self managed byte[] ) is not viable as it breaks type safety and can cause the whole OS to crash ( note there were MANY bugs in Linux here)
* Singularity uses a shared heap and ensures ( via run time checks) there is only a single pointer into the data owned by a single owner( and the shared data has no references to and from other memory)
* Using copy on send for IPC and treating shared memory separately.

# STRUCT MESSAGEs (1)

The system will support small struct messages . These struct messages must be entirely made of unboxed value types ( or BufferManage) and implement ISystemMessage. It is very useful for small calls not requiring strings. Note while you can send a message with char[] you cannot send a message with string and it is not recommended you send message with char[] unless it is very small and necessary.

These structs must not be boxed so the caller cannot use the interface to refer to them as it will cause it to be boxed. The compiler will handle this case specially ie structs sent to SendMessage (IMessage msg) are NOT boxed instead the code to copy them to receiver is inserted by the compiler.

When an unboxed struct message is sent it is copied and added to the receivers GC and hence boxed. Note this is a very fast form of IPC and usefull for quick ping pong messages

# REFERENCE PASSING (2)

For STP’s sharing the same address space (and hence Garbage Collector) STP can directly pass references to other STP’s. This is fast and obviously desired but required that both STPs use the same GC. This normally occurs within a STP trees or on smaller systems. Thread Safety is still required.

# IMMUTABLE DATA

Methods 3 and 4 rely on sending Immutable objects controlled by the GC of the sender. The references are sent to the destination STP they are immutable and hence so no locks are required on the object.

All such objects are IImmutable and they have some special properties

* It’s legal to have references to such objects outside of your address space (which is not the case for other user generated types).
* Objects so marked are checked at compile time to ensure they have no public fields or set properties. Method return types (including ref and out) must also be immutable objects.
* They are treated as fixed by the GC once the reference is sent to another STP.

A lib contains a number of these objects for common usage such as ImmutableArray<T> and ImmutableString. Even though strings are immutable it’s better to use a separate type for IPC.

The system ensures these objects to avoid strange bugs.

Objects marked with the attribute are checked at compile time to ensure they have no public fields or set properties. Method return types (including ref and out) must also be immutable objects.

A TCB system lib will also allow a factory construction such as Immutable.Create<GenType>(params);

# SHARING IMMUTABLE DATA (3)

Data Exchange sends data as sub trees of data. The immutable data trees are created via the constructor of the objects and need to be created leaf first in order. This is the best way to send data between un-trusted sources.

These objects will be created in a shared heap , the compiler will insert ref counting when it is referenced. This count is stored in the object ( which will be set to 0 by the send) .

When a process is shut down or killed reference counts to any such objects must be removed or a memory leak will result.

To ensure objects are created on a SharedHeap a using region is used during which the compiler will change all alloc calls to the Shared Heap. If an object in the GC is referenced by the mutable object a compile error must be generated.

# Other METHODS CoNSIDEReD

1.The following was considered and removed due to complexity and code weight though it could prove effective. Sending mutable data and Buffer manager take precedence.

SHARING NON MUTABLE OBJECTS

Sending Objects with methods such as eg BufferReader . In this case an owner of data can let other STPes act on data via an object provided by the owner. A good example is a BufferManager service may allow streaming reads of a buffer via a BufferReader object. The object needs to ensure all data returned via methods or events are immutable even though the Buffer does not need to be. The object also needs to ensure thread safety if it is manipulating data. Such an object allows the reading or writing of data with no dispatch.

Note many Data Exchange objects have a method that tells the caller they are finished with the data in which case the dummy reference will be released..

Methods in such an object must be careful as they are modifying data via a remote thread and GC . [Consider a using block to change GC]

GC interaction.

When sending a reference the reference is automatically cached locally this “dummy” is needed so the GC does not collect the object.

When a STP is killed or exits all counterparties (i.e. STP which can be reached via an IPC channel) must be sweeped by the kernel and references to object in the address space of the finishing STP must be replaced with null. If in the sweep a dummy object exists than the STP chain needs to be followed. (This handles the issue of a reference being sent through multiple STPed).

The last condition that must be met is that when the receiver signals that it is finished with the object there must be no other references ,this can be achieved by a variety of means. An object can’t be released if it has been sent to another STP (it’s in the proxy) . To prevent this we can do one of the following.

Do a sweep similar to STP termination setting all references to null. These may be bunched up and can be synched with the receiver GC mark phase.

To assist in these things and to avoid the GC being full of fixed objects Immutable object will be created in a separate buffer ( one for each Pipe) sub allocated from the main Memory Manager to avoid issues with fixed objects in the GC. This buffer will reclaim regions of the buffer and will try to keep allocations at similar speed to a Mark Sweep Nursery. It is important to note that this buffer will remain in use until the receiver’s release these objects , if they do not it will additional memory to be consumed by the Pipe as the whole region of the buffer cannot be released until all parts are.

A piece of trusted code is used by the receiving GC to tell the original GC that it is finished with an IPC message. The receiver will then set all references to this object to null and then remove the dummy reference of the sender ( directly ).

SHARING LARGE MUTABLE DATA

2. It is unknown if the following will be needed as the use of immutable data and BufferManager may be sufficient. If 2 processes share a lot of data its prob better to share a GC.

Shared memory managed through the Shared Memory System. It allows changes to a large amount of data between 2 SIAs.

All such classes must be marked with SharedMemory attributes. Memory allocation is via a shared heap. To use this shared memory a pointer is simply passed to a STP after being wrapped in a control block.. If a STP terminates the state of the memory should be fine if it did not hold a lock, if the owner dies it becomes invalid.

All pointers to shared memory are indirect via a trusted control block and will deny access if the user is not the current owner ie there is only 1 user who has visibility of the shared memory at one time. The pointer is contained in a reference to a control block (this is the indirect pointer) This is the recommended method it is secure and reliable. When the lock is not owned the indirect pointer will automatically return null (with the control block hiding the real reference) . A write barrier or compiler check is used to ensure all references are through a control block and references outside of the shared memory region will throw an exception.

A control block exposes an EventWaitHandle , a method for creating new objects , a lock and the immutable reference to the root object in shared memory.

Control blocks are privileged objects. As the only valid reference is through a control block the only way of gaining access is to obtain a block from the system IPC (which requires the share object in its constructor) and/or cloning. The data structure is only removed from memory when all control blocks have been returned. The current lock owner may also release the memory (setting all pointers to control blocks null)

# BUFFER MANAGER

Creating a Buffer Manager requires a Capability. It’s a special piece of trusted code that acts as Immutable and can be referenced from immutable objects ( even a struct Message) . It users pointers internally and allows windows to byte[] data to be passed around while underneath there is a much large buffer. Once Buffer Manager creates a buffer it is fairly static to remove such a buffer would require walking all threads and GCs for references and nulling them. Another example of its use is for a network ring buffer where the driver can choose to expose the Mac Header/IP Header/TCP Header or body to the receiver or a reduced set.

# uPGRADING

Its important that SIAs upgrade to just sending references when they share the same GC . This is quite trivial , however applications written to share GC must always do so and will not work unless we introduce one of the other methods.

# MESSAGE PiPE (QUEUES)

To allow Asynchronous messaging the system uses fixed length queues containing references to the message. The Queues are point to point and in some cases there may not be a return queue as acknowledgements can be called via methods on the message or not required.

Messages queues have a number of benefits namely

* Messages do not need to contain the Sender information
* Security is decoupled from the messaging system ( It is coupled with the queues)
* Backward API Compatibility is trivial to maintain and API changes do not require a recompile. ( especially using inheritance)
* Choice of low latency ping pong or high latency but more cache efficient batch processing
* Better multi Processor support.

It should be noted if the receiver is waiting for the message it gets activated when the message is placed in the queue this allows rapid ping pong IPC for latency critical work.

The receiver makes the reference available to the sender and the sender adds the items. Please note the MessageQueue objects themselves are in the TCB and use pointers to manage the queue.

When a STP starts it normally has a number of queues these are:

* IPC & Events In/Out
* Memory Manager In/Out
* Scheduling In/Out

# MESSAGEs

Message are strongly typed and implement the interface they may inherit from IPC Message or IPCRepplyMessage . Most sub systems will implement their own Sub Class eg MMMessage . Message may be validated. The Message id is equal to the memory address which is valid and unique for the life of the message since it is fixed.

Every STP that does IPC( ie sends and receives messages) has a message pump. Worker STPs working on memory do not (they can be controlled by a parent STP or by waiting for something and it being set in the scheduler)

# EVENTS

The system contains an extensive Publish/ subscribe events system (with security on the subscription). Events work especially well structs and user could even configure call backs on their objects.

Modern systems already have a large number of events namely power management and UI. Extending these seems natural and allows for greater system flexibility without any security cost.

The greatest strength of events is the loose coupling and the low performance overhead for multiple subscribers ( as compared to polling for changes) .

Some of these categories are.

* GC Events (About to collect etc)
* Power Events
* Memory Events
* UI Events
* HW Events ( eg interrupts , USB bus connections )
* Scheduling Signals

One these messages are received they are processed and generate CLR style events. Note messages are only sent to processes that subscribe

# SenD MESSAGES

Note no need for a Syscall for IPC or send message.

Sending STP

Create Immutable object ( or send if it already exists)

queue.SendMessageAsynch(IImuttable) ; // or use the Synch blocking version

Receiving STP

Var message queue.RemoveNextMessage(); // blocking until message available used ManualResetEvent underneath.

STPMessage(message)

// TCB trusted code part of IPC system.

SendMessageAsych (KernelMessage message)

{

bool yield = false;

if ( multiCore == false)

yield = true;

StoreIPCDummyPointer(message); // store dummy so message is not

If (receiver is blocked waiting for message) // issue with receiver resuming after call so need to grab a lock

{

Set Receiver current message to message

Schedule Receiver to run ( via a private reference to the scheduler)

}

else

{

if receiver running

{

Lock (queue) // real solution can void lock

{

if ( high priority)

Add message to front of destination queue)

else

Add message to back of destination queue

}

}

}

if ( yield)

Sleep(0); // give receiver chance to run on single core

return;

} //SendMessage

Receive call return the pipe that has the new message ( use an event ? too expensive ? )

Many to one pipes ? Need to secure each message = bad. So how do we handle multiple blocks // messages by each pipe setting a IPC flag.

IncommingMessage

Reciever

Receive( WaitingOn x