*Todo list

cite this . . . 5
ref this . . . 5
ref image and source . . . 6
ref this . . . 6
update software for ipv6 . . . 6
ref packet headers below . . . 6
ref this . . . 8
Examples of hardware comparisons . . . 13
better word for derive . . . 13
why I need benchmarking . . . 13
show example of optimising java and how different it looks, also different depending on architecture . . . 13
better phrasing off that sentence needed . . . 13
More here . . . 13
get paper . . . 14
explain difference with GC etc . . . 14
get data . . . 14
reference this . . . 15
ref . . . 16
finish this section . . . 16
article on this . . . 16
reasons why JNI is slow . . . 16
ref this . . . 16
ref this . . . 17
expand on this . . . 17
Is this relevant . . . 17
Diagram of packets from NIC using c through 'technique' and then processing packets in java and then back . . . 17
ref this . . . 18
ref this . . . 18
Explain why it has faster performance . . . 19
Nested padding in struct? . . . 19
C struct field always in given order . . . 19
Inconsistencies with datatype length so using uint32t etc . . . 19
Mention this is on 64-bit machine and obviously you don't notice padding . . . 19
Example making sure compiler doesn't pad . . . 19
Proof on speed . . . 20

Imperial College London

Individual Project

Computing - BEng

Packet Processing in Java

Ashley Hemingway

Supervisor:
Peter Pietzuch

2nd Marker:
Wayne Luk

June 2015

Abstract

Acknowledgements

Contents

1 Introduction **4**
- 1.1 Motivation . . . 4
- 1.2 Objectives . . . 4
- 1.3 Solution Idea . . . 4

2 Background **5**
- 2.1 Network Components . . . 5
  - 2.1.1 Network Models . . . 5
  - 2.1.2 Network Packets . . . 6
  - 2.1.3 Packet Handling . . . 9
  - 2.1.4 Network Interface Card (NIC) . . . 9
  - 2.1.5 Network Address Translator (NAT) . . . 9
  - 2.1.6 Firewall . . . 9
- 2.2 Java Features . . . 9
  - 2.2.1 Java Virtual Machine (JVM) . . . 9
  - 2.2.2 Java Native Interface (JNI) . . . 10
  - 2.2.3 Current Java Networking Methods . . . 13
    - 2.2.3.1 Java Sockets . . . 13
    - 2.2.3.2 Remote Method Invocation (RMI) . . . 13
    - 2.2.3.3 Shared Memory Programming . . . 13
- 2.3 jVerbs . . . 13
- 2.4 Native I/O API's . . . 13
  - 2.4.1 Date Plane Development Kit (DPDK) . . . 13
- 2.5 Benchmarking . . . 13
  - 2.5.1 Programming Languages . . . 13
    - 2.5.1.1 Loop Recognition . . . 14
    - 2.5.1.2 Benchmark Game . . . 14
    - 2.5.1.3 Using Economics . . . 14
    - 2.5.1.4 Which is better? . . . 14
  - 2.5.2 Intra-Language Techniques . . . 14
  - 2.5.3 Applications . . . 14

3 DPDK **15**
- 3.1 Environment Abstraction Layer (EAL) . . . 15
- 3.2 huge pages . . . 15
- 3.3 Ring Buffer . . . 15
- 3.4 Memory usage - malloc, mempool, mbuf . . . 15
- 3.5 Poll Mode Driver (PMD) . . . 16

4 Initial language comparison **16**
- 4.1 Benchmarking Algorithm . . . 16
- 4.2 Results . . . 17
- 4.3 Further Investigation . . . 17

5 Design Considerations **17**
5.1 Data Sharing . . . 17
5.1.1 Objects and JNI - using heap and lots of jni calls . . . 18
5.1.2 ByteBuffers - Non-heap and heap memory . . . 18
5.1.3 Java Unsafe - non-heap . . . 18
5.1.4 Evaluation . . . 19
5.1.5 JNA? . . . 19
5.1.6 Packing C Structs . . . 19
5.1.7 Javolution . . . 20
5.1.8 Performance testing . . . 20
5.1.9 Using Packet class directly with c struct data . . . 20

6 Firewall Implementation **20**
6.1 Shared Libraries . . . 20
6.2 Handling Errors . . . 20

7 Evaluation **21**

8 Conclusion **22**

9 User Guide **23**

1 Introduction

1.1 Motivation

1.2 Objectives

1.3 Solution Idea

2 Background

2.1 Network Components

2.1.1 Network Models

Generally there are 2 well known network models. The Open System Interconnection (OSI) model represents an ideal to which all network communication should adhere to while the Transmission Control Protocol/Inter Protocol (TCP/IP) model represents reality in the world. The TCP/IP model combines multiple OSI layers into 1 TCP/IP layer simply because not all systems will go through these exact stages depending on the use of the system.

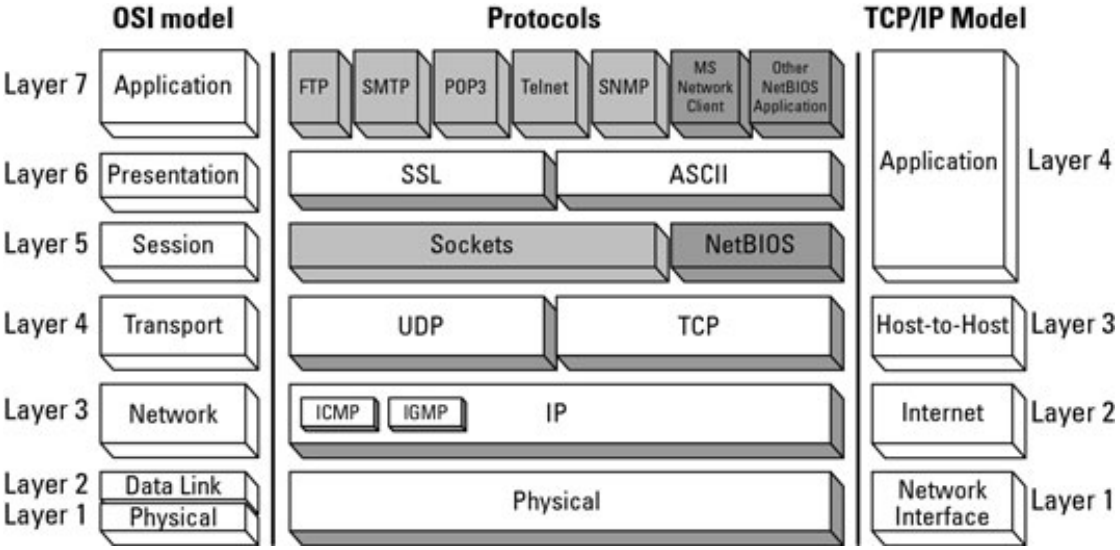

Figure 1: OSI vc TCP/IP Model

cite this

The TCP/IP application layer doesn't represent an actual application but instead its a set of protocols which provides services to applications. These services include HTTP, FTP, SMTP, POP3 and more. It acts as an interface for software to communicate between systems (e.g. client retrieving data from server via SMTP).

The transport layer is responsible for fragmenting the data into transmission control protocol (TCP) or user datagram protocol (UDP) packets, although other protocols can be used. This layer will attach its own TCP or UDP header to the data which contains information such as source and destination ports, sequence number and acknowledgement data.

The network/internet layer attaches a protocol header for packet addressing and routing. Most commonly this will be an IPv4 or IPv6 header . This layer only provides datagram networking functionality and it's up to the transport layer to handle the packets correctly.

ref this

The network interface or link layer will firstly attach its own ethernet header (or suitable protocol header) to the packets, along with an ethernet trailer. This header will specify the destination and source of the media access control (MAC) address which are specific to network interfaces.

The next step is to put the packet onto the physical layer, which may be fibre optic, wireless or standard cables.

This will eventually build a packet of data which include the original raw data along with multiple headers for each layer of the model.

ref image and source

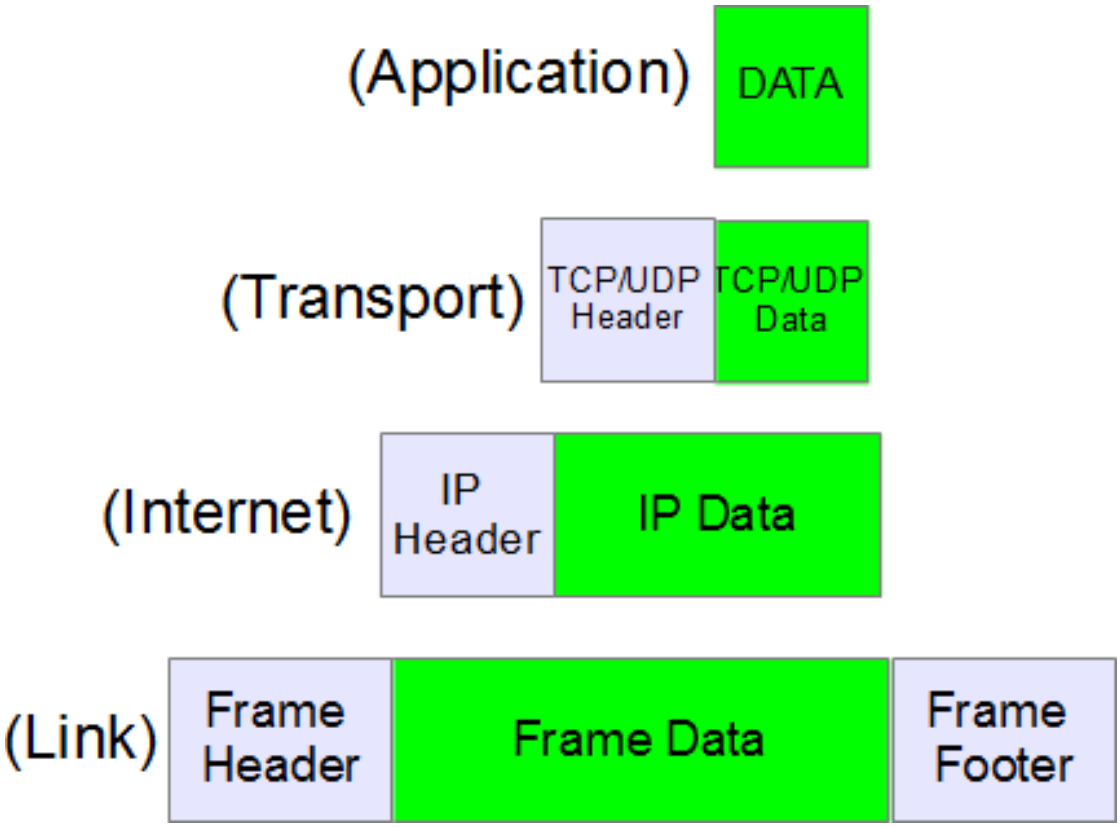

Figure 2: Network model headers [**?**]

2.1.2 Network Packets

A network packet is responsible for carrying data from a source to a destination. Packets are routed, fragmented and dropped via information stored within the packet's header. Note: in this report packets and datagrams are interchangeable. Data within the packets are generally input from the application layer, and headers are appended to the front of this data depending on the network level described in . Packets are routed to their destination based on a combination of an IP address and MAC address which corresponds to a specific computer located within the network, whether that is a public or private network. In this project we will only be concerned with the Internet Protocol (IP) and therefore IPv4 and IPv6 packet headers.

ref this

update software for ipv6

ref packet headers below

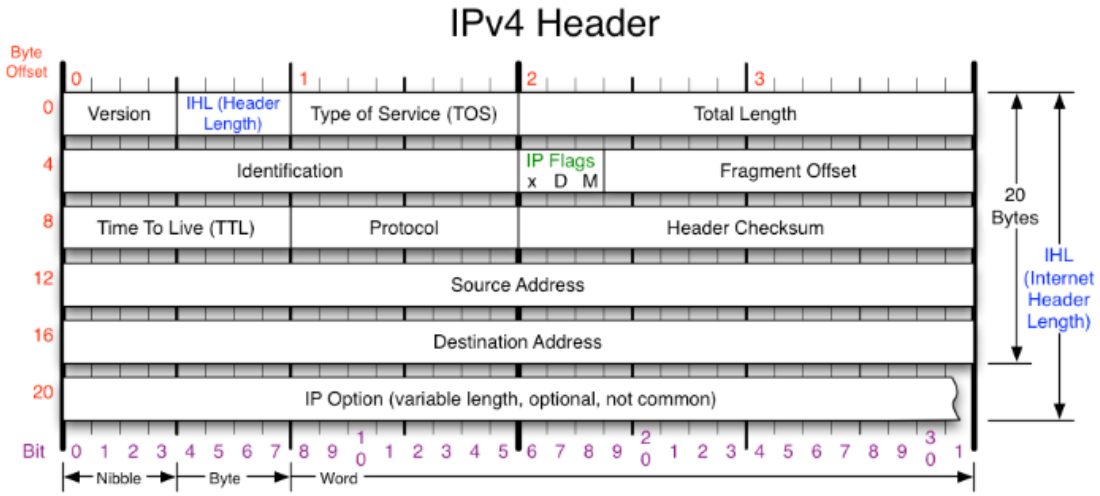

Figure 3: IPv4 Packet Header [?]

- Version - IP version number (set to 4 for IPv4)
- Internet Header Length (IHL) - Specifies the size of the header since a IPv4 header can be of varying length
- Type of Service (TOS) - As of RFC 2474 redefined to be differentiated services code point (DSCP) which is used by real time data streaming services like voice over IP (VoIP) and explicit congestion notification (ECN) which allows end-to-end notification of network congestion without dropping packets
- Total Length - Defines the entire packet size (header + data) in bytes. Min length is 20 bytes and max length is 65,535 bytes, although datagrams may be fragmented.
- Identification - Used for uniquely identifying the group of fragments of a single IP datagram
- X Flag - Reserved, must be zero
- DF Flag - If set, and fragmentation is required to route the packet, then the packet will be dropped. Usually occurs when packet destination doesn't have enough resources to handle incoming packet.
- MF Flag - If packet isn't fragmented, flag is clear. If packet is fragmented and datagram isn't the last fragment of the packet, the flag is set.
- Fragment Offset - Specifies the offset of a particular fragment relative to the beginning of the original unfragmented IP datagram
- Time To Live (TTL) - Limits the datagrams lifetime specified in seconds. In reality, this is actually the hop count which is decremented each time the datagram is routed. This helps to stop circular routing.
- Protocol - Defines the protocol used the data of the datagram

- Header Checksum - Used for to check for errors in the header. Router calculates checksum and compares to this value, discarding if they don't match.
- Source Address - Sender of the packet
- Destination Address - Receiver of the packet
- Options - specifies a number of options which are applicable for each datagram. As this project doesn't concern these it won't be discussed further.

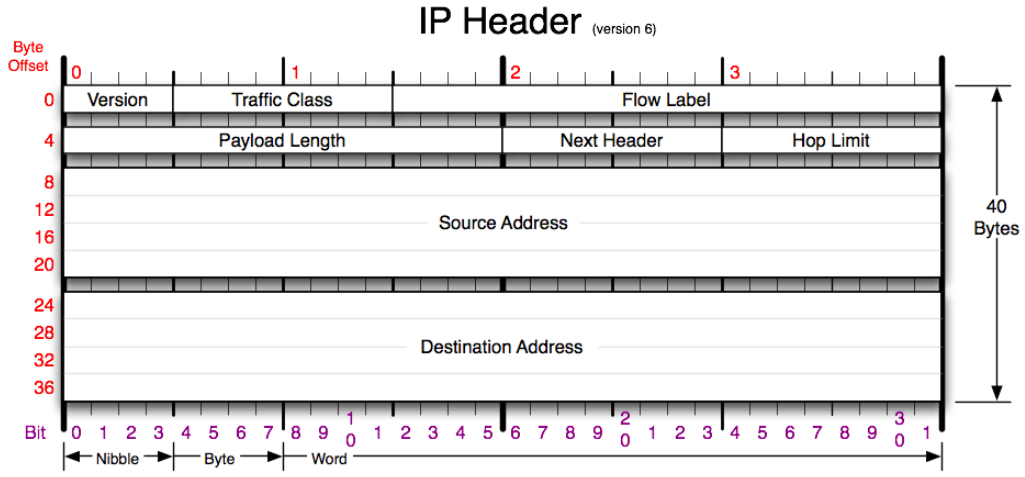

Figure 4: IPv6 Packet Header [?]

- Version - IP version number (set to 6 for IPv6)
- Traffic Class - Used for differentiated services to classify packets and ECN as described in IPv4. ref this
- Flow Label - Used by real-time applications and when set to a non-zero value, it informs routers and switches that these packets should stay on the same path (if multiple paths are available). This is to ensure the packets arrive at destination in the correct order, although other methods for this are available.
- Payload Length - Length of payload following the IPv6 header, including any extension headers. Set to zero when using jumbo payloads for hop-by-hop extensions.
- Next Header - Specifies the transport layer protocol used by packet's payload.
- Hop Limit - Replacement of TTL from IPv4 and uses a hop value decreased by 1 on every hop. When value reaches 0 the packet is discarded.
- Source Address - IPv6 address of sending node
- Destination Address - IPv6 address of destination node(s).
- Extension Headers - IPv6 allows additional internet layer extension headers to be added after the standard IPv6 header. This is to allow more information for features such as fragmentation, authentication and routing information. The transport layer protocol header will then be addressed by this extension header.

2.1.3 Packet Handling

2.1.4 Network Interface Card (NIC)

2.1.5 Network Address Translator (NAT)

2.1.6 Firewall

2.2 Java Features

2.2.1 Java Virtual Machine (JVM)

The Java Virtual Machine is an abstract computer that allows Java programs to run on any computer without dependant compilation. This works by all Java source been compiled down into Java byte code, which is interpreted by the JVM's just in time (JIT) compiler to machine code. However, it does require each computer to have the Java framework installed which is dependant on the OS and architecture. It provides an appealing coding language due to the vast support, frameworks and code optimisations available such as garbage collections and multithreading. Figure 5 shows the basic JVM architecture with the relevant sections explained in the list below.

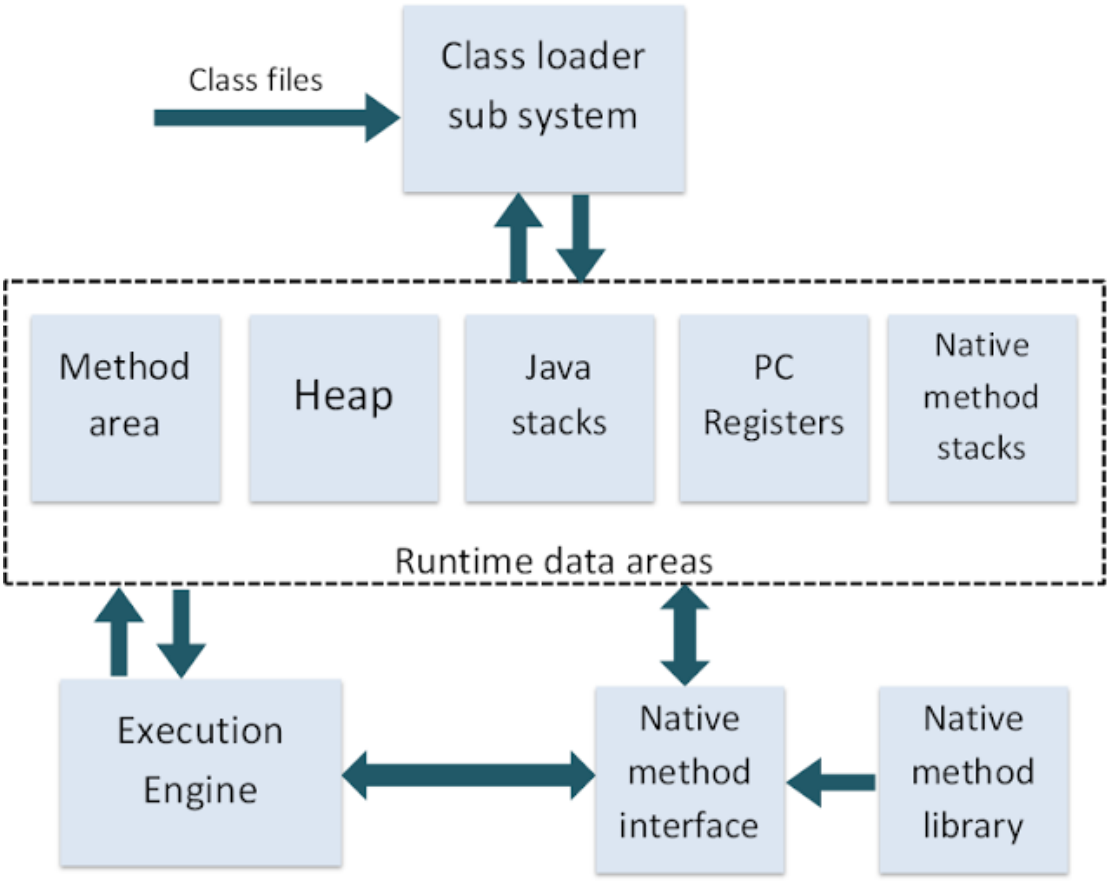

Figure 5: Java Virtual Machine interface [**?**]

- Class loader sub system - Loads .class files into memory, verifies byte code instructions and allocates memory required for the program
- Method area - stores class code and method code
- Heap - New objects are created on the heap
- Stack - Where the methods are executed and contains frames where each frame executes a separate method
- PC registers - Program counter registers store memory address of the instruction to be executed by the micro processor
- Native method stack - Where the native methods are executed.
- Native method interface - A program that connects the native method libraries with the JVM
- Native method library - holds the native libraries information
- Execution engine - Contains the interpreter and (JIT) compiler. JVM decides which parts to be interpreted and when to use JIT compiler.

Typically, any network communication from a Java application occurs via the JVM and through the operating system. This is because the JVM is still technically an application running on top of the OS and therefore doesn't have any superuser access rights. Any network operation results in a kernel system call, which is then put into a priority queue in order to be executed. This is one of the main reasons why network calls through the JVM and kernel can be seen as 'slow', in relative speeds compared to network line rate speeds.

2.2.2 Java Native Interface (JNI)

Provided by the Java Software Development Kit (SDK), the JNI is a native programming interface that lets Java applications use libraries written in other languages. The JNI also includes the invocation API which allows a JVM to be embedded into native applications. This project and therefore this overview will only focus on Java code using C libraries via the JNI on a linux based system.

In order to call native libraries from Java applications a number of steps have to be undertaken as shown below, which are described in more detail later:

1. Java code - load the shared library, declare the native method to be called and call the method
2. Compile Java code - compile the Java code into bytecode
3. Create C header file - the C header file will declare the native method signature and is created automatically via a terminal call
4. Write C code - write the corresponding C source file
5. Create shared library - create a shared library file (.so) from C code
6. Run Java program - run the Java program which calls the native code

The Java Framework provides a number of typedef's used to have equivalent data types between Java and C, such as jchar, jint and jfloat, in order to improve portability. For use with objects, classes and arrays, Java also provides others such as jclass, jobject and jarray so interactions with Java specific characteristics can be undertaken from native code run within the JVM.

```
1  public class Jni {
2
3    int x = 5;
4
5    public int getX() {
6      return x;
7    }
8
9    static { System.loadLibrary("jni"); }
10
11   public static native void objectCopy(Jni o);
12
13   public static void main(String args[]) {
14     Jni jni = new Jni();
15     objectCopy(jni);
16   }
17
18 }
```

Code 1: Basic Java class showing native method declaration and calling with shared library loading

Code 1 shows a simple Java program which uses some native C code from a shared library. Line 9 indicates which shared library to load into the application, which is by default lib*.so where * indicates the name of the library identifier. Line 11 is the native method declaration which specifies the name of the method and the parameters which will be passed to the corresponding C method. In this case, the method name is 'objectCopy' and a 'Jni' object is passed as a parameter. Line 15 is where this native method is called.

```
1 $ javac Jni.java
```

Code 2: Compiling basic Java program

Code 2, run from a terminal, compiles the Java class and create a class file which can be executed.

```
1 $ javah -jni Jni
```

Code 3: Generating C header file

In order to generate the C header file the command 'javah' (Code 3) is used with the flag 'jni' which tells Java that a header file is required which is for use with the JNI. It will then produce method signatures which correspond to the native method declared within Jni.java. The auto generated C header file is shown in Code 4.

```
1  /* DO NOT EDIT THIS FILE - it is machine generated */
2  #include <jni.h>
3  /* Header for class Jni */
4
5  #ifndef _Included_Jni
6  #define _Included_Jni
7  #ifdef __cplusplus
8  extern "C" {
9  #endif
10 /*
11  * Class:     Jni
12  * Method:    objectPrint
```

```
13  * Signature: (LJni;)V
14  */
15 JNIEXPORT void JNICALL Java_Jni_objectPrint
16   (JNIEnv *, jclass, jobject);
17
18 #ifdef __cplusplus
19 }
20 #endif
21 #endif
```

Code 4: Auto-generated C header file

```
1 #include "Jni.h"
2
3 JNIEXPORT void JNICALL Java_Jni_objectPrint(JNIEnv *env, jclass class, jobject obj
      ) {
4   jclass cls = (*env)->FindClass(env, "Jni");
5   jmethodID method = (*env)->GetMethodID(env, cls, "getX", "()I");
6   int i = (*env)->CallIntMethod(env, obj, method);
7   printf("Object x value is %i\n", i);
8 }
```

Code 5: C source file corresponding to auto-generated header file

The C source file implementation is in Code 5. The method signature on line 3 isn't as first declared in the Java source file. The 'env' variable is used to access functions to interact with objects, classes and other Java features. It points to a structure containing all JNI function pointers. Furthermore, the method invocation receives the class from which it was called since it was a static method. If the method had been per instance, this variable would be of type 'jobject'.

Line 4 shows how to find a class identifier by using the class name. In this example, the variables 'class' and 'cls' would actually be equal. In order to call an objects' method, a method id is required as a pointer to this method. Line 5 shows the retrieval of this method id, whose parameters are the jclass variable, the method name and the return type, in this case an integer (represented by an I). Then the method can be called on the object via one of the numerous helper methods (line 6) which differ depending on the return type and static or non-static context.

```
1 $ gcc -shared -fpic -o libjni.so -I/usr/java/include -I/usr/java/include/linux jni
      .c
```

Code 6: Terminal commands to generate shared library file (.so)

The command in Code 6 will create the shared object file called 'libjni.so' from the source file 'jni.c'. This output file is what the Java program uses to find the native code when called. It requires pointers to the location of the Java Framework provided jni.h header file.

```
1 $ java -Djava.library.path="." Jni
2 Object x value is 5
```

Code 7: Output from running Java application calling native C methods

Running the Java application, pointing to the location where Java can find the shared library (if not in a standard location) will output the above in Code 7.

Although the JNI provides a very useful interface to interact with native library code, there are a number of issues that users should be wary of before progressing:

- The Java application that relies on the JNI loses its portability with the JVM as it relies on natively compiled code.

- Errors within the native code can potentially crash the JVM, with certain errors been very difficult to reproduce and debug.
- Anything instantiated with the native code won't be collected by the garbage collector with the JVM, so freeing memory should be a concern.
- If using the JNI on large scale, converting between Java objects and C structs can be difficult

2.2.3 Current Java Networking Methods

For high performance computing in Java, a number of existing programming options are available in order for applications to communicate over a network. These can be classified as: (1) Java sockets; and (2) Remote Method Invocation (RMI); (3) shared memory programming. As will be discussed, none of these are capable of truly high performance networking, especially at line rate speeds.

2.2.3.1 Java Sockets

Java sockets are the standard low level communication for applications as most networking protocols have socket implementations. They allow for streams of data to be sent between applications as a socket is one end point for a 2 way communication link, meaning that data can be read from and written to a socket in order to transfer data. Even though sockets are a viable option for networking, both of the Java socket implementations (IO sockets & NIO (new I/O) sockets) are inefficient over high speed networks [**?**] and therefore lack the performance that is required. As discussed previously, the poor performance is due to the JVM interacting with network cards via the OS kernel.

2.2.3.2 Remote Method Invocation (RMI)

Remote Method Invocation (RMI) is a protocol developed by Java which allows an object running in a JVM to invoke methods on another object running on a different JVM. Although this method provides a relatively easy interface for which JVM's can communicate, its major drawback relates to the speed. Since RMI uses Java sockets as its basic level communication method, it faces the same performance issues as mentioned in section **??**.

2.2.3.3 Shared Memory Programming

Shared memory programming provides high performance JVM interaction due to Java's multithread and parallel programming support. This allows different JVM's to communicate via objects within memory which is shared between the JVM's. However, this technique requires the JVM's to be on the same shared memory system, which is a major drawback for distributed systems as scalability options decrease.

Even though these 3 techniques allow for communication between JVM's and other applications, the major issue is that incoming packets are still handled by the kernel and then passed onto the corresponding JVM. This means that packets are destined for certain applications, meaning that generic packets can't be intercepted and checked, which is a requirement for common middlebox software.

2.3 jVerbs

Ultra-low latency for Java applications has been partially solved by the jVerbs [**?**] framework. Using remote direct memory access (RDMA), jVerbs provides an interface for which Java applications can communicate, mainly useful within large scale data centre applications.

RDMA is a technology that allows computers within a network to transfer data between each other via direct memory operations, without involving the processor, cache or operating system of either communicating computer. RDMA implements a transport protocol directly within the network interface card (NIC), allowing for zero copy networking, which allows a computer to read from another computer and write to its own direct main memory without intermediate copies. High throughput and performance is a feature of RDMA due to the lack of kernel involvement, but the major downside is that it requires specific hardware which supports the RDMA protocol, while also requiring the need for specific computer connections set up by sockets.

As jVerbs takes advantage of mapping the network device directly into the JVM, bypassing both the JVM and operating system (Figure **??**), it can significantly reduce the latency. In order to have low level interaction with the NIC, jVerbs has a very thin layer of JNI calls which can increase the overhead slightly. However, jVerbs is flawed, mainly because it requires specific hardware to run on, firstly limited by the RDMA protocol reliant hardware and further by the required RDMA wrappers which are implemented by the creators. Also, it can only be used for specific computer to computer connection and not generally packet inspection.

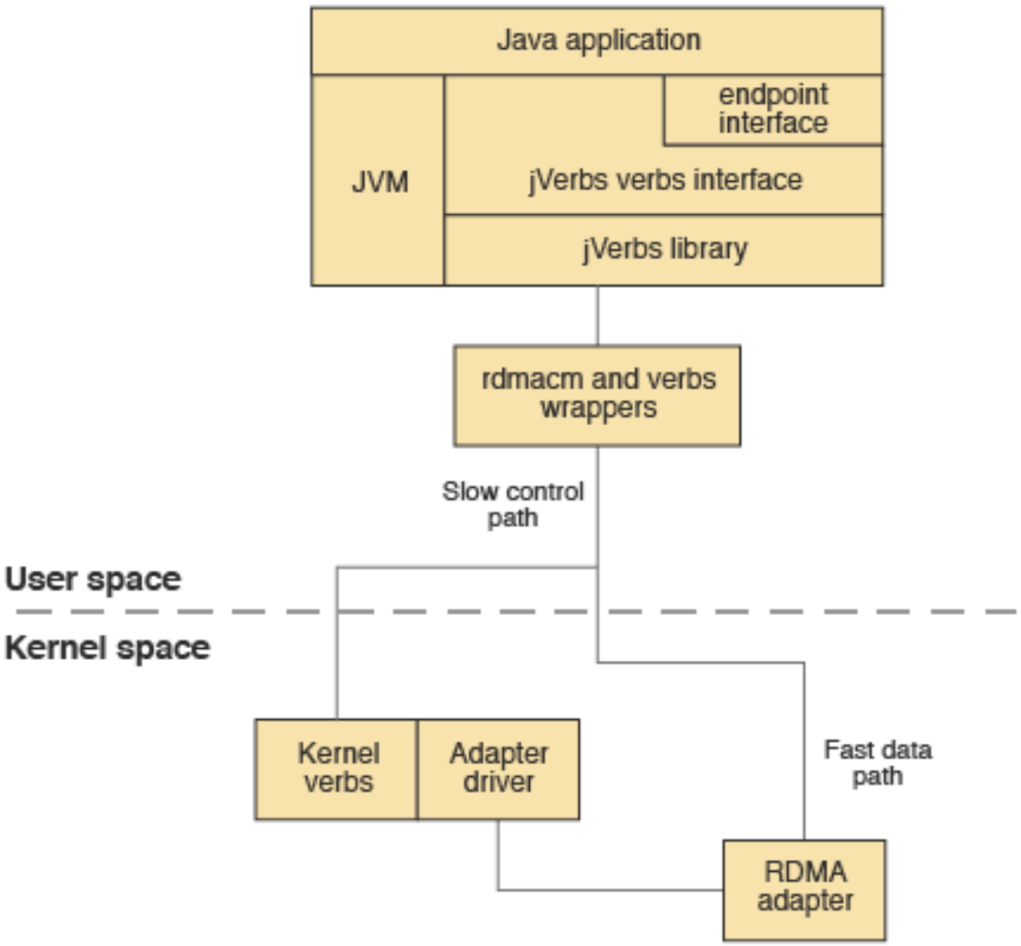

Figure 6: jVerbs architecture - shows how the framework bypasses the kernel and JVM [**?**]

jVerbs provides a useful example framework which re-emphasises that packet processing in Java is very possible with low latency, while assisting in certain implementation and design choices which can be analysed in more detail.

2.4 Native I/O API's

Currently available native networking API's are capable of reading and writing packets to the NIC transmission and receive queues at line rate. This is due to a number of techniques which tend to alter the kernels understand of the underlying NIC, therefore requiring specialist hardware and software to use such tools. DPDK (Section **??**) is one of the tools that is open source and publicly available for use.

2.4.1 Data Plane Development Kit (DPDK)

Data Plane Development Kit (DPDK) [**?**] is a set of libraries and drivers which enabled fast

refer to in-depth section - also is this right?

packet processing with certain system set ups. Since DPDK is developed by Intel, it only supports Intel x86 CPU's and certain network interface controllers (NIC). DPDK overwrites the NIC's drivers meaning that the operating system doesn't recognise the network cards and can't interact with them. It installs its own drivers allowing it to interact with certain memory locations without permission from the kernel or even involving it in any way.

DPDK makes use of an environment abstraction layer (EAL) which hides the environmental specifics and provides a standard interface which any application can interact with. Due to this, if the system changes in any way, the DPDK library needs to be re-compiled with other features been re-enabled in order to allow applications to run correctly again.

In order to use the DPDK libraries for the intended purpose, data packets have to be written into the correct buffer location so they are inserted onto the network. I similar approach is used when receiving packets on the incoming buffer ring, but instead of the system using interrupts to acknowledge the arrival of a new packet, which is performance costly, it constantly polls the buffer space to check for new packets. DPDK also allows for multiple queues per NIC and can handle multiple NIC's per system, therefore scalability is a major bonus of the libraries.

DPDK is very well documented on a number of levels. Firstly there is a online API which gives in depth details about what the methods, constants and structs do. There are a number of well written guides which give step-by-step details of how to install, set-up and use DPDK on various platforms and finally, there are many sample programs included with the build which give understanding of how the overall library works.

2.5 Benchmarking

Benchmarking is a process of testing hardware, individual components or full end to end systems to determine the performance of the application or hardware . Generally, benchmarking should be repeatable under numerous iterations without only minor variations in performance results. This is firstly to allow minor changes to be made to the application/component with re-runs of the benchmark showing the performance changes. Secondly, it allows accurate comparisons to be drawn between similar software or hardware with different implementations in order to derive a better product.

Examples of hardware comparisons

better word for derive

why I need benchmarking

2.5.1 Programming Languages

It is well known that different programming languages can provide a radical change in execution for a given program. However, direct comparisons can't truly be trusted as certain languages are suited for for specific tasks and finding a benchmarking program to incorporate this is problematic. Other factors can be introduced when deciding on the optimisation level and the compiler of JIT used.

show example of optimising java and how different it looks, also different depending on architecture

Numerous attempts have been made to compare languages, most noticeably the 'Benchmark Game' and Google.

2.5.1.1 Loop Recognition

Google inducted their own experiment on this problem, testing only C++, Java, Scala and Go on

better phrasing off that sentence needed

More here

the loop recognition algorithm . Implementations made use of standard looping constructs and memory allocation schemes without the use of non-orthodox optimisation techniques. Selected results of this are shown below:

get paper

explain difference with GC etc

Benchmark	Time [sec]	Facter
c++	23	1.0x
Java 64-bit	134	5.8x
Java 32-bit	290	12.8x
Java 32-bit GC	106	4.6x
Scala	82	3.6x
Go 6g	161	7.0x

Table 1: Results from Loop Recognition benchmarking

2.5.1.2 Benchmark Game

The Benchmark Game is an online community which aims to find the best programming language by using multiple benchmarking algorithms running on different architecture configurations to determine the outcome. Again, even this community regard the best benchmark application to be your application. A few selected results are shown below between Java and C (those used in this report) for a few different benchmarks.

get data

2.5.1.3 Using Economics

2.5.1.4 Which is better?

2.5.2 Intra-Language Techniques

2.5.3 Applications

3 DPDK

This section will go into much more detail about the Data Plane Development Kit (DPDK), focussing on the basic concepts used by the fast packet processing framework for use within custom applications.

3.1 Environment Abstraction Layer (EAL)

The EAL abstracts the specific environment from the user and provides a constant interface in order for applications to be ported to other environments without problems. The EAL may change depending on the architecture (32-bit or 64-bit), operating system, compilers in use or network interface cards. This is done via the creation of a set of libraries that are environment specific and are responsible for the low-level operations gaining access to hardware and memory resources.

On the start-up of an application, the EAL is responsible for finding PCI information about devices and addresses via the igb_uio kernel module, performing physical memory allocation using huge pages and starting user-level threads for the logical cores (lcores) of the application. reference this

3.2 huge pages

Huge pages are a way of increasing performance when dealing with large amounts of memory. Normally, memory is managed in 4096 byte blocks known as pages, which are listed within the CPU memory management unit (MMU). However, if a system uses a large amount of memory, increasing the number of standard pages is expensive on the CPU as the MMU can only handle thousands of pages and not millions.

The solution is to increase the page size from 4KB to 2MB or 1GB (if supported) which keeps the number of pages references small in the MMU but increases the overall memory for the system. Huge pages should be assigned at boot time for better overall management, but can be manual assigned on initial boot up if required.

DPDK makes uses of huge pages simply for increased performance due to the large amount of packets in memory. This is even the case if the system memory size is relatively small and the application isn't processing an extreme number of packets.

3.3 Ring Buffer

A ring buffer is used by DPDK to manage transmit and receive queues for each network port on the system. This fixed sized first-in-first-out system allows multiple objects to be enqueued and dequeued from the ring at the same time from any number of consumers or producers. A major disadvantage of this is that once the ring is full (more of a concern in receive queues), it no more objects can be added to the ring, resulting in dropped packets. It is therefore imperative that applications can processes packets at the required rate.

3.4 Memory usage - malloc, mempool, mbuf

DPDK can make use of non-uniform memory access (NUMA) if the system supports it. NUMA is a method for speeding up memory access when multiple processors are trying to access the same memory and therefore reduces the processors waiting time. Each processor will receive its own bank of memory which is faster to access as it doesn't have to wait. As applications become

more extensive, processors may need to share memory, which is possible via moving the data between memory banks. This somewhat negates the need for NUMA, but NUMA can be very effective depending on the application. DPDK can make extensive use of NUMA as each logical core is generally responsible for its own queues, and since queues can't be shared between logical cores, data sharing is rare.

As discussed previously , DPDK uses hugepages in memory and therefore it provides its own malloc library, which as expected, allocates hugepage memory to the user. However, as DPDK focusses on raw speed, the use of the DPDK malloc isn't suggested as pool-bases allocation is faster. DPDK also provides ways to malloc memory on specific sockets depending on the NUMA configuration of the application.

ref

finish this section

3.5 Poll Mode Driver (PMD)

A Poll Mode Driver (PMD) is responsible

4 Initial language comparison

Before any implementation or specific design considerations were undertaken, an evaluation of the performance of C, Java and Java using the Java Native Interface (JNI) was carried out. Although data from existing articles and websites could be used for Java and C, there was no existing direct comparisons between them and the JNI, therefore custom tests were carried out.

The JNI is inherently seen as a bottleneck of an application (even after its vast update in Java 7).

article on this

As this application would be forced to use the JNI, numeric values of its performance was helpful to evaluate the bridge in speed required to be overcome.

reasons why JNI is slow

4.1 Benchmarking Algorithm

As discussed previously , there are always advantages and disadvantages of any algorithm used for benchmarking. In order to minimise the disadvantages, an algorithm was used which tried to mimic the procedures which would be used in the real application, just without the complications. Algorithm 1 shows that the program basically creates 100,000 packets individually and populates their fields with random data, which is then processed and return in the 'result' field. This simulates retrieving low-level packet data, interpreting and acting upon the data, and then setting data within the raw packet.

ref this

Algorithm 1 Language Benchmark Algorithm

```
1:  function Main
2:      for i = 1 to 100000 do
3:          p ← Initialise Packet
4:          PopPacket(p)
5:          ProPacket(p)

6:  function PopPacket(Packet p)                    ▷ Set data in a packet
7:      p.a ← randomInt()
8:      p.b ← randomInt()
9:      p.c ← randomInt()
10:     p.d ← randomInt()
11:     p.e ← randomInt()

12: function ProPacket(Packet p)                    ▷ Process a packet
13:     res ← p.a * p.b * p.c * p.d * p.e
14:     p.result ← res
```

For the JNI version, the same algorithm was used, however, the PopPacket method was carried out on the native side to simulate retrieving raw packet data. The ProPacket method was executed on the Java side with the result been passed back to the native side.

4.2 Results

Each language had the algorithm run 1000 times in order to minimise any variations due to external factors. Figures show that C was considerably quicker than Java, while Java using the JNI was extremely slow.

ref this

expand on this

4.3 Further Investigation

Due to the very poor performance of the JNI compared to other languages, further investigations were carried out to find more specific results surrounding the JNI.

Is this relevant

5 Design Considerations

5.1 Data Sharing

The proposed application will be sharing data between the DPDK code written in C and the Java side used for the high functionality part of the application. This requires a large amount of data, most noticeably packets, to be transferred between 'sides' in a small amount of time.

Diagram of packets from NIC using c through 'technique' and then processing packets in java and then back

A few techniques for this are available with Java and C, all with different performances and ease-of-use.

5.1.1 Objects and JNI - using heap and lots of jni calls

By far the simplest technique available is using the Java Native Interface (JNI) in order to interact with native code and then retrieve the required via this. This can be done 2 ways, either by creating the object and passing it as a parameter to the native methods or creating an object on the native side via the Java environment parameter. Both ways require the population of the fields to be done on the native side. From then on, any data manipulation and processing could be done on the Java side. Unfortunately, this does require all data to be taken from the object and placed back into the structs before packets can be forwarded. Obviously this results in a lot of unneeded data copying, while the actual JNI calls can significantly reduce the speed of the application as shown in .

ref this

5.1.2 ByteBuffers - Non-heap and heap memory

ByteBuffers are a Java class which allow for memory to be allocated on the Java heap (non direct) or outside of the JVM (direct). Non direct ByteBuffer's are simply a wrapper for a byte array on the heap and are generally used as they allow easier access to bytes, as well as other primitive data types.

Direct ByteBuffers allocate memory outside of the JVM in native memory. This firstly means that the only limit on the size of ByteBuffers is memory itself. Furthermore, the Java garbage collector doesn't have access to this memory. Direct ByteBuffers have increased performance since the JVM isn't slowed down by the garbage collector and intrinsic native methods can be called on the memory for faster data access.

5.1.3 Java Unsafe - non-heap

The Java Unsafe class is actually only used internally by Java for its memory management. It generally shouldn't be used within Java since it makes a safe programming language like Java an unsafe language (hence the name) since memory access exceptions can be thrown. It can be used for a number of things such as:

- Object initialisation skipping
- Intentional memory corruption
- Nullifying unwanted objects
- Multiple inheritance
- Dynamic classes
- Very fast serialization

Obviously without proper precautions any of these actions can be dangerous and can result in crashing the full JVM. This is why the Unsafe class has a private constructor and calling the static Unsage.getUnsafe() will throw a security exception for untrusted code which is hard to bypass. Fortunately, Unsafe has its own instance called 'theUnsafe' which can be accessed by using Java reflection :

ref this

```
1 Field f = Unsafe.class.getDeclaredField("theUnsafe");
2 f.setAccessible(true);
3 Unsafe unsafe = (Unsafe) f.get(null);
```

Code 8: Accessing Java Unsafe

Using Unsafe then allows direct native memory access to retrieve data in any of the primitive data formats. Custom objects with a set structure can then be created, accessed and altered using Unsafe which provides a vast increase in performance over traditional objects stored on the heap. This is mainly thanks to the JIT compiler which can use machine code more efficiently.

5.1.4 Evaluation

5.1.5 JNA?

5.1.6 Packing C Structs

Structs are a way of defining complex data into a grouped set in order to make this data easier to access and reference as shown in Code 9.

```
struct example {
    char *p;
    char c;
    long x;
    char y[50];
    int z;
};
```

Code 9: Example C Struct

On modern processors all commercially available C compilers will arrange basic C datatypes in a constrained order to make memory access faster. This has 2 effects on the program. Firstly, all structs will actually have a memory size larger than the combined size of the datatypes in the struct as a result of padding. However, this generally is a benefit to most consumers as this memory alignment results in a faster performance when accessing the data.

Explain why it has faster performance

Nested padding in struct?

C struct field always in given order

Inconsistencies with datatype length so using uint32t etc

Code 10 shows a struct which has compiler inserted padding. Any user wouldn't know the padding was there and wouldn't be able to access the data in the bits of the padding through conventional C dereferencing paradigm. This example does assume use on a 64-bit machine with 8 byte alignment, but 32-bit machines or a different compiler may have different alignment rules.

```
struct example {
    char *p;        // 8 bytes
    char c;         // 1 byte
    char pad[7];    // 7 byte padding
    short x;        // 2 bytes
    char pad[6];    // 6 byte padding
    char y[50];     // 50 bytes
    int z;          // 4 bytes
};
```

Code 10: Example C Struct with compiler inserted padding

Mention this is on 64-bit machine and obviously you don't notice padding

Example making sure compiler doesn't pad

Since the proposed application in this report requires high throughput of data, the initial thought would be that this optimisation is a benefit to the program. Generally this is the case, but for data which is likely to be shared between the C side and Java side a large amount, data accessing is far quicker on the Java side if the struct is padded. This results in certain structs been forced to be padded when compiled, more noticeably, those used for packet headers.

Proof on speed

5.1.7 Javolution

5.1.8 Performance testing

5.1.9 Using Packet class directly with c struct data

6 Firewall Implementation

6.1 Shared Libraries

6.2 Handling Errors

7 Evaluation

8 Conclusion

9 User Guide