1 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.

1.1 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

1.2 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.

1.3 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.

1.4 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`

1.5   Poll Mode Driver (PMD)

A Poll Mode Driver (PMD) is responsible

2   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`

2.1   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:         $p \leftarrow$ *Initialise Packet*
4:         POPPACKET(p)
5:         PROPACKET(p)

6: **function** POPPACKET(Packet p)   ▷ Set data in a packet
7:     $p.a \leftarrow$ *randomInt()*
8:     $p.b \leftarrow$ *randomInt()*
9:     $p.c \leftarrow$ *randomInt()*
10:     $p.d \leftarrow$ *randomInt()*
11:     $p.e \leftarrow$ *randomInt()*

12: **function** PROPACKET(Packet p)   ▷ Process a packet
13:     $res \leftarrow p.a * p.b * p.c * p.d * p.e$
14:     $p.result \leftarrow res$

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.

2.2 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

2.3 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

3 Design Considerations

3.1 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.

3.1.1 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 .

3.1.2 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.

3.1.3 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 :

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

3.1.4 Evaluation

3.1.5 JNA?

3.1.6 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 2.

```
struct example {
    char *p;
    char c;
    long x;
    char y[50];
    int z;
};
```
Code 2: 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 4 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 (only via pointer arithmetic). 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 3: Example C Struct with compiler inserted padding

> Mention this is on 64-bit machine and obviously you don't notice padding and order of elements can pay an important part in this

5

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 packed (no padding). This results in certain structs been forced to be packed when compiled, more noticeably, those used for packet headers.

Proof on speed

Packed structures mean there are no gaps between elements, required alignment is set to 1 byte. Also __attribute__((packed)) definition means that compiler will deal with accessing members which may get misaligned due to 1 byte alignment and packing so reading and writing is correct. However, compilers will only deal with this misalignment if structs are accessed via direct access. Using a pointer to a packed struct member (and therefore pointer arithmetic) can result in the wrong value for the dereferenced pointer. This is since certain members may not be aligned to 1 byte. In the below example, unint32 is 4 byte aligned and therefore it is possible for a pointer to it to expect 4 byte alignment therefore resulting in the wrong results.

```c
#include <stdio.h>
#include <inttypes.h>
#include <arpa/inet.h>

struct packet {
    uint8_t x;
    uint32_t y;
} __attribute__((packed));

int main ()
{
    uint8_t bytes[5] = {1, 0, 0, 0, 2};
    struct packet *p = (struct packet *)bytes;

    // compiler handles misalignment because it knows that
    // "struct packet" is packed
    printf("y=%"PRIX32", ", ntohl(p->y));

    // compiler does not handle misalignment - py does not inherit
    // the packed attribute
    uint32_t *py = &p->y;
    printf("*py=%"PRIX32"\n", ntohl(*py));
    return 0;
}
```

Code 4: Example C Struct with compiler inserted padding

On an x86 system (which does not enforce memory access alignment), this will produce
y=2, *py=2
as expected. On the other hand on my ARM Linux board, for example, it produced the seemingly wrong result
y=2, *py=1
However, since a packed struct is much easier to traverse from Java than a padded struct, the decision was made to make certain structs packed within the DPDK framework and then recompile the libraries. This decision could be made since other structs within the DPDK framework were also packed and therefore consideration of this was already made.

Note that if a struct contains another struct, that struct should be packed recursively as-well to ensure the first struct has no padding at all.

Char doesn't have alignment and can start on any address. But 2-byte shorts must start on an even address, 4-byte ints or floats must start on an address divisible by 4, and 8-byte longs or doubles must start on an address divisible by 8. Signed or unsigned makes no difference.

Self-alignment makes access faster because it facilitates generating single-instruction fetches and puts of the typed data. Without alignment constraints, on the other hand, the code might end up having to do two or more accesses spanning machine-word boundaries. Characters are a special case; theyre equally expensive from anywhere they live inside a single machine word. Thats why they dont have a preferred alignment.

casting to an odd pointer will slow down code and could work. Other architectures will take the word which the pointer points to and therefore the problem occurs above.

3.1.7 Javolution

3.1.8 Performance testing

3.1.9 Thread affinity

Thread.currentThread().getId(); just gets id of thread relative to jvm.

It keeps a process limited to certain a certain core or cores. Process will still be taken out of use and switched back in but without the problem of moving cache between cores.

Normally as a thread gets a time slice (a period in which to use the core), it is granted whichever core [CPU] is determined to be most free by the operating system's scheduler. Yes, this is in contrast to the popular fallacy that the single thread would stay on a single core. This means that the actual thread(s) of an application might get swapped around to non-overclocked cores, and even underclocked cores in some cases. As you can see, changing the affinity and forcing a single-threaded CPU to stay on a single CPU makes a big difference in such scenarios. The scaling up of a core does not happen instantly, not by a long shot in CPU time.
Therefore, for primarily single (or limited) thread applications, it is sometimes best to set the CPU affinity to a specific core, or subset of cores. This will allow the 'Turbo' processor frequency scaling to kick in and be sustained (instead of skipping around to various cores that may not be scaled up, and could even be scaled down).

core thrashing - ust by the name, you know this is a bad thing. You lose performance when a thread is swapped to a different core, due to the CPU cache being 'lost' each time. In general, the *least* switching of cores the better. One would hope the OS would try to avoid this, but it doesn't seem to at all in quick tests under Windows 7. Therefore, it is recommended you manually adjust the CPU affinity of certain applications to achieve better performance.

Another important issue is avoiding placing a load on a HyperThreaded (non-physical) core. These cores offer a small fraction of the performance of a real core. The Windows scheduler is aware of this and will swap to them only if needed. As of mid Jan 2012 the Windows 7 and Windows 2008 R2 schedulers have a hotfix for AMD Bulldozer CPUs that see them as Hyper-Threaded, cutting them down from 8 physical cores to 4 physical cores, 8 logical cores. This is for two reasons: The AMD Bulldozer platform uses pairs of cores called Bulldozer Modules. Each pair shares some computation units, such as an L2 cache and FPU. To spread out the load

and prevent too much load being placed on two cores that have shared computational units, the Windows patch was released, boosting performance in lightly threaded scenarios.

Processor affinity takes advantage of the fact that some remnants of a process that was run on a given processor may remain in that processor's memory state (for example, data in the CPU cache) after another process is run on that CPU. Scheduling that process to execute on the same processor could result in an efficient use of process by reducing performance-degrading situations such as cache misses. A practical example of processor affinity is executing multiple instances of a non-threaded application, such as some graphics-rendering software.

put code showing cpu affinity assignment

In Linux, Java thread uses the native thread(i.e, thread provided by Linux). This means the JVM creates a new native thread when the Java code creates a new java thread. So, the Java threads can be organised in any way the native threads can be organised.

A native thread can be bound to a core through the sched_setaffinity() function. So, a Java thread can be bound to a core. If Java standard library does not provide a function to do so, then this function need to be provided through JNI.

In Linux, multi-threading is same as parallel threading. Linux kernel distribute threads among processors to balance the cpu load. However the individual threads can be bound with any core as wished. So, in Linux Java multi-threading is same as parallel threading.

4 General Implementation

4.1 Shared Libraries

4.2 Handling Errors