Advanced Operating Systems: Lab 2 - IPC

Dr Robert N. M. Watson

2020-2021

The goals of this lab are to:

- Continue to gain experience tracing user-kernel interactions via system calls and traps.
- Explore the implementations and performance of varying IPC models and buffering approaches.
- Use DTrace and hardware performance counters (HWPMC) to analyse these properties.
- Generate data to complete the second lab assignment.

You will do this by using DTrace and HWPMC to analyse the behaviour of a potted, kernel-intensive IPC benchmark.

Background: POSIX IPC objects

POSIX defines several types of Inter-Process Communication (IPC) objects, including pipes (created using the `pipe()` system call) and sockets (created using the `socket()` and `socketpair()` system calls).

Pipes are used most frequently between pairs of processes in a UNIX *process pipeline*: a chain of processes started by a single command line, whose output and input file descriptors are linked. Although pipes can be set up between unrelated processes, the primary means of acquiring a pipe is through inheritance across `fork()`, meaning that they are used between closely related processes (e.g., with a common parent process).

Sockets are used when two processes are created in independent contexts and must later rendezvous – e.g., via the filesystem, but also via TCP/IP. In typical use, each endpoint process creates a socket via the `socket()` system call, which are then interconnected through use of `bind()`, `listen()`, `connect()`, and `accept()`. However, there is also a `socketpair()` system call that returns a pair of interconnected endpoints in the same style as `pipe()` – convenient for us as we wish to compare the two side-by-side.

Both pipes and sockets can be used to transmit ordered byte streams: a sequence of bytes sent via one file descriptor that will be received reliably on the other without loss or reordering. As file I/O, the `read()` and `write()` system calls can be used to read and write data on file descriptors for pipes and sockets. It is useful to know that these system calls are permitted to return *partial reads* and *partial writes*: i.e., a buffer of some size (e.g., 1k) might be passed as an argument, but only a subset of the requested bytes may be received or sent, with the actual size returned via the system call's return value. This may happen if the in-kernel buffers for the IPC object are too small for the full amount, or if *non-blocking I/O* is enabled. When analysing traces of IPC behaviour, it is important to consider both the size of the buffer passed and the number of bytes returned in evaluating the behaviour of the system call.

You may wish to read the FreeBSD `pipe(2)` and `socketpair(2)` manual pages to learn more about these APIs before proceeding with the lab. These are installed on your RPi board, and can be read using the commands `man 2 pipe` and `man 2 socketpair`.

The benchmark

As with our earlier I/O benchmark, the IPC benchmark is straightforward: it sets up a pair of IPC endpoints referencing a shared pipe or socket, and then performs a series of `write()` and `read()` system calls on the file descriptors to send (and then receive) a total number of bytes of data. Data will be sent using a smaller userspace buffer size – although as hinted above, there is no guarantee that a full user buffer will be sent or received in any individual call. Also as with the I/O benchmark, there are several modes of operation: sending and receiving within a single thread, a pair of threads in the same process, or between two threads in two different processes.

The benchmark will set up any necessary IPC objects, threads, and processes, sample the start time using the `clock_gettime()` system call, perform the IPC loop (perhaps split over two threads), and then sample the finish time using the `clock_gettime()` system call. Optionally, both the average bandwidth across the IPC object, and also more verbose information about the benchmark configuration, may be displayed. A single, dynamically linked binary will be used in this lab: `ipc-benchmark`.

The UNIX command line

All commands will be run as the root user. Example command lines are prefixed with the `#` symbol signifying the shell prompt; you should type in only text after the prompt.

Obtaining the benchmark

You can find the benchmark source code on your RPi in:

`/advopsys-packages/labs/2020-2021-advopsys-lab2.tbz`.

We recommend untarring this file into the `/data` directory on your board:

```
# cd /data ; tar -xzf /advopsys-packages/labs/2020-2021-advopsys-lab2.tbz
```

Compiling the benchmark

The laboratory IPC benchmark source code has been preinstalled onto your RPi4 board. However, you will need to build it before you can begin work. Once you have logged into your RPi4 (see *Advanced Operating Systems: Lab Setup*), build the bundle:

```
# make -C ipc
```

Running the benchmark

Once built, you can run the benchmark binary as follows, with command-line arguments specifying various benchmark parameters:

```
# ipc/ipc-benchmark
```

If you run the benchmark without arguments, a small usage statement will be printed, which will also identify the default IPC object type, IPC buffer, and total IPC sizes configured for the benchmark. As in the prior lab, you will wish to be careful to hold most variables constant in order to isolate the effects of specific variables. For example, you might wish the vary the IPC object type while holding the total IPC size constant.

In addition to a set of arguments specifying parameters for the benchmark itself, which will feel familiar from the prior lab, there is a new argument to request that hardware performance counters be measured around the benchmark run.

Required operation argument

While this benchmark supports multiple modes of operation, this lab will use only one mode:

2thread Run the benchmark between two threads within one process: one as a 'sender' and the other as a 'receiver', with the sender capturing the first timestamp, and the receiver capturing the second. System calls are blocking, meaning that if the in-kernel buffer fills during a `write()`, then the sender thread will sleep; if the in-kernel buffer empties during a `read()`, then the receiver thread will sleep.

Optional I/O flags

-b *buffersize* Specify an alternative userspace IPC buffer size in bytes – the amount of memory allocated to hold to-be-sent or received IPC data. The same buffer size will be used for both sending and receiving. The total IPC size must be a multiple of buffer size.

-g Collect `getrusage()` statistics, such as sampled user and system time, as well as message send and receive statistics.

-i *ipctype* Specify the IPC object to use in the benchmark: `pipe`, `local`, or `tcp` (default `pipe`).

-j Generate output as JSON, allowing it to be more easily imported into the Jupyter Lab framework, as well as other data-processing tools.

-P *mode* Enable performance counters across the IPC loop. See the document, *Advanced Operating System: Hardware Performance Counters (HWPMC)* for information on the available modes and their interpretation.

-s When operating on a socket, explicitly set the in-kernel socket-buffer size to match the userspace IPC buffer size rather than using the kernel default. Note that per-process resource limits will prevent use of very large buffer sizes.

-t *totalsize* Specify an alternative total IPC size in bytes. The total IPC size must be a multiple of userspace IPC buffer size.

Terminal output flags

The following arguments control terminal output from the benchmark; remember that output can substantially change the performance of the system under test, and you should ensure that output is either entirely suppressed during tracing and benchmarking, or that tracing and benchmarking only occurs during a period of program execution unaffected by terminal I/O:

-q *Quiet mode* suppress all terminal output from the benchmark, which is preferred when performing whole-program benchmarking.

-v *Verbose mode* causes the benchmark to print additional information, such as the time measurement, buffer size, and total IPC size.

Example benchmark commands

This command performs a simple IP benchmark using a pipe and default userspace IPC buffer and total IPC sizes between two threads of the same process:

```
# ipc/ipc-benchmark -i pipe 2thread
```

This command performs a socket-pair benchmark, and requests non-default socket-buffer sizes synchronised to a userspace IPC buffer size of 1k:

```
# ipc/ipc-benchmark -i local -s -b 1024 2thread
```

As with the I/O benchmark, additional information can be requested using *verbose mode*:

```
# ipc/ipc-benchmark -v -i pipe 2thread
```

And, likewise, all output can be suppressed, and *bare mode* can be used, for whole-program analysis:

```
# ipc/ipc-benchmark -q -B -i pipe 2thread
```

This command instructs the IPC benchmark to capture information on memory instructions issued when operating on a socket with a 512-byte buffer from a single thread:

```
# ipc/ipc-benchmark -i local -b 512 -P tlbmem 1thread
```

This command performs the same benchmark while tracking L1 data-cache and L2 cache hits and refills:

```
# ipc/ipc-benchmark -i local -b 512 -P dcache 1thread
```

This command performs the same benchmark while tracking architectural loads, stores, function returns, and exception returns:

```
# ipc/ipc-benchmark -i local -b 512 -P arch 1thread
```

Note on kernel configuration

By default, the kernel limits the maximum per-socket socket-buffer size that can be configured, in order to avoid resource starvation. You will need to tune the kernel's default limits using the following command, run as root, prior to running benchmarks. Note that this should be set before any benchmarks are run, whether or not they are explicitly configuring the socket-buffer size, as the limit will also affect socket-buffer auto-sizing.

```
# sysctl kern.ipc.maxsockbuf=33554432
```

Notes on using DTrace

On the whole, this lab will be concerned with just measuring the IPC loop, rather than whole-program behaviour. As in the last lab, it is useful to know that the system call `clock_gettime` is both run immediately before, and immediately after, the IPC loop. In this benchmark, these events may occur in different threads or processes, as the sender performs the initial timestamp before transmitting the first byte over IPC, and the receiver performs the final timestamp after receiving the last byte over IPC. You may wish to bracket tracing between a return probe for the former, and an entry probe for the latter; see the notes from the last lab for an example.

As with the last lab, you will want to trace the key system calls of the benchmark: `read()` and `write()`. For example, it may be sensible to inspect `quantize()` results for both the execution time distributions of the system calls, and the amount of data returned by each (via `arg0` in the system-call return probe). You will also want to investigate scheduling events using the `sched` provider. This provider instruments a variety of scheduling-related behaviours, but it may be of particular use to instrument its `on-cpu` and `off-cpu` events, which reflect threads starting and stopping execution on a CPU. You can also instrument `sleep` and `wakeup` probes to trace where threads go to sleep waiting for new data in an empty kernel buffer (or for space to place new data in a full buffer). When tracing scheduling, it is useful to inspect both the process ID (`pid`) and thread ID (`tid`) to understand where events are taking place.

By its very nature, the probe effect is hard to investigate, as the probe effect does, of course, affect investigation of the effect itself! However, one simple way to approach the problem is to analyse the results of performance benchmarking with and without DTrace scripts running. When exploring the probe effect, it is important to consider not just the impact on bandwidth average/variance, but also on systemic behaviour: for example, when performing more detailed tracing, causing the runtime of the benchmark to increase, does the number of context switches increase, or the distribution of `read()` return values? In general, our interest will be in the overhead of probes rather than the overhead of terminal I/O from the DTrace process – you may wish to suppress that output during the benchmark run so that you can focus on probe overhead.

Notes on benchmark

As with the prior lab, it is important to run benchmarks more than once to collect a distribution of values, allowing variance to be analysed. You may wish to discard the first result in a set of benchmark runs as the system will not yet have entered its steady state. Do be sure that terminal I/O from the benchmark is not included in tracing or time measurements (unless that is the intent).

Note on graphs in this lab assignment or lab report

Because of the large amounts of data (and number of data sets) explored in this lab, you will need to pay significant attention in writing your lab assignment or lab report to how you present data visually. Graphs should make visual arguments, and how a set of graphs are plotted can support (or confuse) that argument. Make sure all graphs are clearly presented with labels and textual descriptions helping the reader identify the points you think are important.

When two graphs have the same independent variable (e.g., buffer size), it is important that they use the same X axis in terms of labelling and scale. Graphs with the same X axis will often benefit from being arranged so that they align horizontally on the page, such that inflection points can be visually compared.

Where an X axis is identical, and dependent variables have the same Y axis (e.g., both measure bandwidth and have the same scale), placing them on the same graph is frequently useful, as visual artefacts (such as intersecting lines, differing slopes) have specific meaning and will pop out at readers. Be careful to clearly label different lines, and ideally use shading, point symbol, and/or colour to make the visual distinction clear. If you have having trouble distinguishing the different data sets, then there are too many data sets on the graph.

Where an X axis is identical, but dependent variables differ on their scales (e.g., one measures bandwidth, and a second cache refill rate), placing them on the same graph could lead to confusion as, for example, line intersections may not actually have meaning. You can, however, vertically stack multiple graphs on the same X axis, allowing inflection points and changes in slopes to be visually compared. Do this by aligning the X axes of the two graphs, and then 'squinching' (a technical term) the two close together; as the X axes will have identical units and values, you can have the graphing package include labels only for the bottom graph. This will allow comparison of linked data – e.g., a larger graph showing bandwidth, and a set of smaller graphs showing micro-architectural effects such as TLB and cache refill rates, to be visually compared to make it easy to assess possible correlation.