

IHK/McKernel Tutorial

Balazs Gerofi

RIKEN Advanced Institute for Computational Science,
JAPAN

2015/Dec/14



Agenda

- **Loading IHK modules**
- **Configuring the IHK resource pool**
- **Creating an OS instance**
- **Assigning resources to the OS instance**
- **Loading kernel image and setting kernel arguments**
- **Booting the kernel and inspecting kernel logs**
- **Running a simple directory list**
- **Running NAS parallel benchmarks**
- **Destroying the OS instance**
- **Unloading IHK modules**
- **Adding a new system call**
- **Recompiling IHK and McKernel**
- **Calling the syscall from userspace**

Load IHK modules, check IHK device file, list CPUs

```
hpcuser@cn01:~$ cd install/3.19.0-25-generic/  
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo insmod kmod/ihk.ko  
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo insmod kmod/ihk-smp-x86.ko  
hpcuser@cn01:~/install/3.19.0-25-generic$  
hpcuser@cn01:~/install/3.19.0-25-generic$ ls -ls /dev/mcd0  
0 crw----- 1 root root 250, 0 Dec 11 19:06 /dev/mcd0  
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo chown hpcuser /dev/mcd0  
hpcuser@cn01:~/install/3.19.0-25-generic$  
hpcuser@cn01:~/install/3.19.0-25-generic$ lscpu | grep CPU  
CPU op-mode(s):      32-bit, 64-bit  
CPU(s):              4  
On-line CPU(s) list: 0-3  
CPU family:          6  
CPU MHz:             2300.000  
NUMA node0 CPU(s):   0-3
```

Configure IHK resource pool (CPU and memory)

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 reserve cpu 2,3
hpcuser@cn01:~/install/3.19.0-25-generic$
hpcuser@cn01:~/install/3.19.0-25-generic$ lscpu | grep CPU
CPU op-mode(s):      32-bit, 64-bit
CPU(s):              4
On-line CPU(s) list: 0,1
Off-line CPU(s) list: 2,3
CPU family:          6
CPU MHz:             2300.000
NUMA node0 CPU(s):   0,1

hpcuser@cn01:~/install/3.19.0-25-generic$ cat /proc/meminfo | grep MemFree
MemFree:             1359392 kB
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 reserve mem 512M@0
hpcuser@cn01:~/install/3.19.0-25-generic$ cat /proc/meminfo | grep MemFree
MemFree:             835112 kB
```

Create a lightweight co-kernel instance

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 create
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo chown hpcuser /dev/mcos0
hpcuser@cn01:~/install/3.19.0-25-generic$ ls -ls /dev/mcos0
0 crw----- 1 hpcuser root 249, 0 Dec 11 19:28 /dev/mcos0
```

Assign resources to OS instance

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 assign cpu 2,3
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 assign mem 512M@0

hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 query cpu
2-3

hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 query mem
536870912@0
```

Load kernel image and set kernel arguments

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 load smp-x86/kernel/  
mckernel.img
```

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 kargs "hidos"
```

Boot kernel and load syscall offload support module

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 boot
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo insmod kmod/mcctrl.ko
hpcuser@cn01:~/install/3.19.0-25-generic$
hpcuser@cn01:~/install/3.19.0-25-generic$ dmesg | tail -n 10
[ 1101.799647] Connect msg: 25, 20000001f5, 0, 1587a000
[ 1101.799650] syscall: MC CPU 1 connected. c=ffff88005a9c9000
[ 1101.799651] (Accepted) Remote channel id = 0
[ 1101.799652] os(ffff880059efe000)->mchannel = ffff88005acc7c00
[ 1101.799713] IKC init: cpu=0 port=502
[ 1101.799720] Request: 5a790000, Response: 15859000, Doorbell: 3699b000
[ 1101.799721] Request: ffff88005a790000, Response: ffff880015859000,
Doorbell: ffff88003699b000
[ 1101.799742] IKC init: cpu=1 port=501
[ 1101.799749] Request: 5ac60000, Response: 15880000, Doorbell: 3699b000
[ 1101.799750] Request: ffff88005ac60000, Response: ffff880015880000,
Doorbell: ffff88003699b000
```

Display McKernel kernel log

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 kmsg
MCK started.
[ -1]: ns_per_tsc: 434
[ -1]: CPU: 2 3
[ -1]: map_start = 15800000, map_end = 35800000
[ -1]: TEXT: # of large pages = 2
[ -1]: TEXT: Base address = 15a00000
[ -1]: Page table is now at ffff800015a4b000
[ -1]: no_execute_available: 1
[ -1]: map_fixed: fee00000 => ffffffff70000000 (1 pages)
[ -1]: Invariant TSC supported.
[ -1]: setup_x86 done.
[ -1]: map_fixed: 59410000 => ffffffff70001000 (1 pages)
[ -1]: KCommand Line: hidoss
[ -1]: Page allocator: 15800000 - 35800000 (12)
[ -1]: Available pages: 130984 pages
[ -1]: Page allocator: fffff00000000000 - fffff00100000000 (12)
[ -1]: map_fixed: 8a000 => ffffffff70002000 (2 pages)
[ -1]: Trampoline area: 0x8a000
[ -1]: map_fixed: 0 => ffffffff70004000 (1 pages)
[ -1]: # of cpus : 2
[ -1]: locals = ffff800015822000
[ -1]: IKC IRQ vector: 64, IKC target CPU APIC: 0
[ 0]: BSP HW ID = 2
[ 0]: AP Booting: 1 (HW ID: 3)
[ 0]: AP Booting: Done
[ 0]: Master channel init acked.
MCK/IHK booted.
```

Run a simple directory list

```
hpcuser@cn01:~/install/3.19.0-25-generic$ bin/mcexec ls -ls
total 16
4 drwxr-xr-x 2 hpcuser hpcuser 4096 Nov 14 10:17 bin
4 drwxr-xr-x 2 hpcuser hpcuser 4096 Nov 14 10:17 kmod
4 drwxr-xr-x 2 hpcuser hpcuser 4096 Nov 14 10:17 sbin
4 drwxrwxr-x 4 hpcuser hpcuser 4096 Nov 11 02:29 smp-x86
```

Run NAS parallel benchmarks CG

```
hpcuser@cn01:~/install/3.19.0-25-generic$ bin/mcexec ~/src/NPB3.3.1/NPB3.3-OMP/bin/cg.A.x
```

NAS Parallel Benchmarks (NPB3.3-OMP) - CG Benchmark

Size: 14000

Iterations: 15

Number of available threads: 2

Initialization time = 0.278 seconds

iteration	r	zeta
1	0.25633607125478E-12	19.9997581277040

. . . /* output stripped */

Please send all errors/feedbacks to:

NPB Development Team

npb@nas.nasa.gov

```
hpcuser@cn01:~/install/3.19.0-25-generic$
```

Destroy OS instance and release resources

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 destroy 0
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo rmmod mcctrl.ko
hpcuser@cn01:~/install/3.19.0-25-generic$
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 release cpu 2,3
hpcuser@cn01:~/install/3.19.0-25-generic$ lscpu | grep CPU
CPU op-mode(s):      32-bit, 64-bit
CPU(s):              4
On-line CPU(s) list: 0-3
CPU family:          6
CPU MHz:             2300.000
NUMA node0 CPU(s):   0-3
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkconfig 0 release mem 512M@0
hpcuser@cn01:~/install/3.19.0-25-generic$ cat /proc/meminfo | grep MemFree
MemFree:             1342284 kB
```

Adding a new system call (1/2)

- Got to ~/src/mckernel
- Edit “arch/x86/kernel/include/syscall_list.h”
- Add new entry to syscall table:

```
diff --git a/arch/x86/kernel/include/syscall_list.h b/arch/x86/kernel/include/syscall_list.h
index 6e2eef9..0f29aa0 100644
--- a/arch/x86/kernel/include/syscall_list.h
+++ b/arch/x86/kernel/include/syscall_list.h
@@ -138,6 +138,7 @@ @@ SYSCALL_HANDLED(303, mod_call)
  SYSCALL_HANDLED(309, getcpu)
  SYSCALL_HANDLED(310, process_vm_readv)
  SYSCALL_HANDLED(311, process_vm_writev)
+SYSCALL_HANDLED(599, kprint)
  SYSCALL_HANDLED(601, pmc_init)
  SYSCALL_HANDLED(602, pmc_start)
  SYSCALL_HANDLED(603, pmc_stop)
```

Adding a new system call (2/2)

- Edit “kernel/syscall.c”
- Add implementation of kprint call:

```
diff --git a/kernel/syscall.c b/kernel/syscall.c
index 37f21b2..f5c7680 100644
--- a/kernel/syscall.c
+++ b/kernel/syscall.c
@@ -6127,6 +6127,14 @@ static void do_mod_exit(int status){
 }
 #endif

+SYSCALL_DECLARE(kprint)
+{
+    char *user_buf = (char *)ihk_mc_syscall_arg0(ctx);
+    kprintf("%s\n", user_buf);
+    return 0;
+}
+
/* select counter type */
SYSCALL_DECLARE(pmc_init)
{
```

Recompiling IHK and McKernel

- Create build directories, run configure scripts and make

```
hpcuser@cn01:~$ mkdir -p ~/build/ihk+mckernel/ihk
hpcuser@cn01:~$ mkdir -p ~/build/ihk+mckernel/mckernel
hpcuser@cn01:~$ cd ~/build/ihk+mckernel/ihk/
hpcuser@cn01:~/build/ihk+mckernel/ihk$ ~/src/ihk/configure --prefix=/home/hpcuser/install/`uname -r`/ --
with-target=smp-x86

. . .

hpcuser@cn01:~/build/ihk+mckernel/ihk$ cd ../mckernel/
hpcuser@cn01:~/build/ihk+mckernel/mckernel$ ~/src/mckernel/configure --prefix=/home/hpcuser/install/
`uname -r`/ --with-target=smp-x86

. . .

hpcuser@cn01:~/build/ihk+mckernel/mckernel$ cd ../ihk/
hpcuser@cn01:~/build/ihk+mckernel/ihk$ make && make install

. . .

hpcuser@cn01:~/build/ihk+mckernel/ihk$ cd ../mckernel/
hpcuser@cn01:~/build/ihk+mckernel/mckernel$ make && make install

. . .

hpcuser@cn01:~/build/ihk+mckernel/mckernel$ ls ~/install/`uname -r`/
bin  kmod  sbin  smp-x86
```

Write user-space app that calls the new syscall

- `vim ~/src/kprint_user.c`

```
#define _GNU_SOURCE
#include <stdio.h>
#include <unistd.h>
#include <sys/syscall.h>

int main(int argc, char **argv)
{
    char buf[1024];
    sprintf(buf, "%s", "Hello Kernel Space!\n");

    syscall(599, buf);

    return 0;
}
```

- **Compile application:**

```
hpcuser@cn01:~/src$ gcc -Wall kprint_user.c -o kprint_user
```

Reboot, run and check results

- Reboot McKernel (either with steps before or using reboot script) and run application:

```
hpcuser@cn01:~/src$ cd ~/install/`uname -r`
hpcuser@cn01:~/install/3.19.0-25-generic$ sudo sbin/mcreboot.sh
hpcuser@cn01:~/install/3.19.0-25-generic$

hpcuser@cn01:~/install/3.19.0-25-generic$ bin/mcexec ~/src/kprint_user
hpcuser@cn01:~/install/3.19.0-25-generic$
```

- Check kernel log:

```
hpcuser@cn01:~/install/3.19.0-25-generic$ sbin/ihkosctl 0 kmsg | tail -n 10
[ -1]: locals = ffff800015822000
[ -1]: IKC IRQ vector: 64, IKC target CPU APIC: 0
[  0]: BSP HW ID = 2
[  0]: AP Booting: 1 (HW ID: 3)
[  0]: AP Booting: Done
[  0]: Master channel init acked.
MCK/IHK booted.
[  0]: Hello Kernel Space!
```

Thank you for your attention!
Questions?