3 MK7.3 Known Problems

3.1 Introduction

This section describes known problems remaining in the release at distribution. Problems are listed by affected area: Kernel, Server, Commands, Libraries, or Unidentified (if the particular area has not been determined or does not apply). The report number for the problem is presented in square brackets at the end of the description (i.e., [ ]).

All known problems are listed here, regardless of which release (microkernel-only or full microkernel/server) is described by these Release Notes.

3.2 Kernel

- DIPC loopback tests do not pass when KKT/Eth is being used. The kernel panics when an rpc local to a node is attempted. [2413]
- The kernel task swapper uses poor criteria for choosing tasks to swap out. It can swap out the only running processes in the system as configured. This is because it treats as “active”, any task that is either running or waiting. With this algorithm, it can swap out all the truly active tasks, leaving an idle shell swapped in. [2381]
- Panic: Cthread wakeup occurs under a heavy system load. An activation is allowed to run before the cthread wired port is set. [2379]

- The swapout thread attempts to do work as soon as the thread is started (boot time). As a result of this it garbage collects the system even though at boot time the load is light. [2376]
- The getrusage system call page fault statistics are always set to 0. [2362]
- When trying to open a file with a pathname longer than the maximum pathname length, the error EFAULT is returned. It should be returning the error ENAMETOOLONG. [2356]
- Assertion failure in *task_info()* “Assertion failed: file “task.c”, line 1629: thread”. Improperly failing in cases where the thread does not yet have a shuttle. [2350]
- System hangs due to a race between thread creation and task swapping. If a task is marked for swapout while it is creating new threads, the new threads will not be marked for swapout. [2337]
- The ZONE_DEBUG option which is enabled under DEBUG mach kernels doesn’t adjust the maximum size of a zone with the added fields inserted to maintain a queue of allocated zone entries. This makes the zone artificially smaller since it can contain fewer elements for the given maximum size. A simple test that attempts to allocate TASK_MAX tasks under a DEBUG kernel will fail due to this. [2330]
- Panic: ipc_task_init when running mach_perf task_perf test. This is due to the reaper_thread not running often enough to maintain a pool of free tasks. [2329]
- SCSI subsystem can’t be unconfigured in config.devices. The kernel fails to link due to many undefined references. [2314]
- memory_object_set_attributes_common causes a memory leak if MACH_PAGEMAP and NORMA_VM are defined. [2312]
- Vm_fault will sometimes sleep interruptibly. This should not occur since vm_fault can be called with locks held. If the system sleeps interruptibly it can deadlock with threads waiting on the lock held by a suspended or swapped out thread. [2290]
- Unresolvable page fault in kernel space. An ordinary page fault (probably just to restore a displaced mapping) on a server buffer is not being serviced properly. The problem is that the kernel addresses which map this memory have been removed, not just from the pte, but from the kernel map itself. The problem is caused by code in the I/O error return path. In this particular dump, it appears that there were a couple I/O errors which preceded the crash. Inspection of the ds_read_done routine indicates that the an incorrect memory free may occur in the error path for I/O which is targeted to the colocated server address space.
- At least 2 MIG .defs files (mach.defs and mach_host.defs) have too many interfaces. This causes some overlaps in the msgh_id of different

messages. For example, *task_set_port_space()* overlaps with *processor_set_destroy()* and *device_map()*. [2274]

- Panic "aha U/OVRN on scsi command" occurs on system where rboot was written to the disk with dd command without the conv=sync option. [2269]
- Heavy I/O over a serial lines may result in the line hanging. Output may be resumed with a CTRL-C. This occurs when software flow control is enabled and *ttyinput()* discards a CTRL-Q character because the input queue is full. [2226]
- System panics when page frames that should have been initialized in *pmap_startup()* are found to be uninitialized memory when referenced by *vm_page_release()*. [2225]
- When running CORDS udpiptest with all user space protocols over Myrinet with a FAST+RT kernel, about 2 milliseconds of jitter is seen in the elapsed times. No jitter is seen with a FAST kernel, or with protocols in the kernel space. [2222]
- A CORDS +net+test kernel will only work properly if the system is in single user when x_client or x_server is set to 1 to bootstrap the cords protocol stack. This is due to a limitation of the BOOTP protocol, which only works in kernel mode iff it is the first protocol to use the network device. [2221]
- *Scsi_probe()* calls from the NCR 53C810 driver do not detect a DEC RZ26L disk drive from an MK7.3 DEBUG kernel. The drive is detected from a PRODUCTION kernel. An MK7.3 DEBUG kernel can detect the drive if breakpoints are set in the right places, so I suspect the problem is either timing related or an uninitialized pointer. [2132]
- When running a collocated server, virtual addresses given by a Unix process are not checked in copyin & copyout routines. Unix processes can corrupt kernel address space, resulting in panics and hangs. [2119]
- While working with Mach aware gdb the kernel died with a bus error. The apparent cause of the bus error was that *task_by_pid()* called mach_port_deallocate with a "name" parameter of -1. [2114]
- VSX cfset{i,o}speed tests fail at 19200 and 38400 baud. This may be a symptom of the fact that the driver (mach_kernel/i386/AT386/com.c) doesn't make use of any buffering supplied by the chip. [2017]
- Test 2 of the RPC chains and alert tests (rpctest_test_suite.sh) sometimes results in an assertion failure: file "/afs/ri/project/kernel/build/mainline/src/mach_kernel/ipc/ipc_right.c", line 651: port->ip_srights > 0. Test 1 of the same suite hangs on a 133 MHz system, but not at 100 MHz. The rpc_client/rpc_server cases tests sometimes fail with the message "rpc_client: port_array_out: badptr" when started at case "a". [1976]

- There is a vm_priv inversion in the zone package. The package makes sure that no more then one thread expands a zone at a time. If a non vm_priv thread hits a zone needed in the pageout path and tries to expand the zone, and the system is out of memory, we will deadlock. The non vm_priv thread will take the “expanding” token and the vm_priv thread will come in later and block waiting for the non vm_priv threads to complete. It never will. [1886]
- When running mach_perf in multiuser mode with other processes simultaneously (such as build), mach_perf will do a task_terminate which will send dead name notifications to the server from ports belonging to other Unix processes, which are therefore killed by the server with a signal 9. [1870]
- When removing the -k argument from the bootstrap.conf, the OSF/1 server crashes in old_get_config_info because it tries to access bad memory from argv. This is due to the bootstrap_argument MIG interface which gets a zero in the argument pointer field. The size argument is ok, so it might be a out of line memory passing problem. [1869]
- This appears to be another tty/mouse bugs. When typing in a vi xterm window (started from the Reply button of exmh), echoing suddenly became very slow (about 1 second). The mouse was frozen too, although the second hand of xclock was still ticking. The machine could be pinged, but neither rlogin nor telnet was successful (telnet said it connected). CTL-ALT-BKSP was used to kill the X11R6 server. It restored VGA text mode successfully, but the shell on the console was not responding. Broke into the debugger with CTL-ALT-D and found that the system was completely idle. Most threads in the server were stuck in either getblk (due to a failed try_lock) or an AFS function called Lock_Acquire. [1846]
- VSE fails when doing the truncate and ftruncate tests. The tests involves setting a file-size limit and then attempting to truncate a file upward in size beyond the file size limit. The test fails when the server posts a signal 25 for exceeding file size. POSIX says it shouldn’t do this on a truncate. [1843]
- The error message ‘vnode_pager_change_attributes: timed out after ten seconds’ occurs one to several times a day. When it does, a ten second plus delay in interactivity occurs. The error message does not indicate a fatal problem; it just shows that there would have been a deadlock while AFS was trying to reclaim inodes. If it fails, it will try again. A “large” AFS cache configuration will reduce the frequency of the messages. [1832]
- When running with *vm_remap()* disabled, the server uses *vm_write()* and *vm_read_overwrite()* to perform *copyin()*/*copyout()*. Unfortunately, the implementation of these interfaces uses virtual copies, not physi-

cal copies, yet the COW faults are taken 100% of the time, resulting in both a performance pessimization and a virtual space leak. [1018]

- The pmap_create function should single-thread the memory chunk allocator to avoid multiple allocations. This bug is merely wasteful and then only if NCPUS > 1, hence it is at a low priority. [126]
- The fix made for mk6 CR59 to make the short-circuited RPC path honor a MIG_NO_REPLY return from the server will need to be extended once *thread_abort()* works on migrated threads. [120]
- The Adaptec 154x driver is not intelligent about how it copies buffers from the low 16MB of memory to higher memory on EISA-bus machines that can support more than 16MB. Note that this problem will not manifest itself when using Adaptec 174x board's operating in enhanced mode. [96]

3.3 MK Server

- OSF1 server cannot be built without STREAMS. Removing the STREAMS option plus all its suboptions yields numerous undefined symbols. [2395]
- The default_pager, though multi-threaded and configured to use asynchronous writes to the paging partition, winds up running single-threaded and with synchronous write operations. This is because it gives up its mapping to the page before starting an asynchronous write. The unmapping will not complete until the I/O is complete, so the default pager just hangs in the object termination code waiting for paging in progress to go to 0. The effect is to turn an async write into a synchronous write. While it is waiting, the kernel holds the map lock for the address space of the default pager. This has the effect of blocking other default pager threads that wish to allocate or deallocate memory. [2387]
- If a second drive is specified in /etc/fstab for automounting, the system will fail to properly mount that or the primary drive most of the time. An fsck failure is reported on one of the partitions. If it happens to be the root partition, the system is essentially dead. A workaround is to delete the second disk from /etc/fstab and mount it manually after the system is in multiuser mode. [2276]
- With the Celebris GL 5100 integrated audio devices enabled, problems occur using X11R6 and selfhost sup collection three. The XFree86 server appears to hang (blank screen with frozen cursor) when creating certain new windows. The hanging occurred with the audio configured for IRQ 5 and either DMA channel 0 or 1. [2272]
- The osf1_server.FAST+AFS.INK server appears to use the wrong default root device. If /mach_servers/bootstrap.conf does not specify a startup command line option, the single server claims to be using “ma-

jor=5, minor=0" as the root device. If the startup line contains an explicit root device (e.g. "-k startup startup.INK sd0a") then the server uses the correct root device (major=3, minor=0). [2271]

- Pthread mutex deadlocks, uxthread_yield should release the uniproc lock and call thread_switch with SWITCH_OPTION_DEPRESS. [2189]
- Breakpoints set with gdb in a user application at *thread_switch()* stop at the breakpoint but hang when trying to continue from the breakpoint. This is the ptrace version of gdb, rev 4.16. [2186]
- Data files in an AFS file system are sometimes overwritten with blocks of nulls. [2133]
- In an AFS directory, overwriting an existing executable file often results in a file which fails to execute and instead produces a failure such as: "rtld main: Load of "mut" failed: Exec format error". File "mut" had been overwritten after a small change in the source file. If mut is deleted and the same link done again, it executes ok. [1981]
- System hangs with the message "can't map region (error = 3)". What seems to have happened is that a vm_map done in elf_getxfile failed, so the process was killed. When the process called exit, it eventually got to exec_ungetxfile, which noticed that the SPAGV flag was set and tried to release a reference on the text vnode. However, the text vnode was never set up. [1978]
- VSX test c_iflag failures when ISTRIP or BRKINT flag is set. [1955]
- On a system running X, server panics produce a blank screen with no input possible. [1934]
- Call to *sleep()* from a cthreads child never returns. [1914]
- It is possible to mount an NFS filesystem to a mountpoint and then mount a local device on top of it. For example, I mounted a remote machine's /usr partition as /mnt, and then mounted /dev/fd0 on top of it. It is not possible to mount an NFS partition on top of an existing NFS mount point. [1893]
- The uniproc mutex is not release and then taken back when invoking device_read/write in *disk_read/write()*. Thus server activity is blocked when waiting for I/O to complete. [1809]
- When "ps ax" is used, the server calls thread_info to get information about each thread of each process. Sometimes (perhaps if processes are exiting at the same time ps is invoked), it passes a bad port name as the target thread port in thread_info. In the case analyzed, this port was one for which the (kernel-loaded) server was the receiver. This can crash the kernel or server, because the kernel will believe that the port name is valid, as long as it is not null. We have a collocation design defect here, in that dead-name usage by kernel-loaded servers cannot be

detected. We also have a server bug, because our server supposedly never tries to use a dead name, as long as no task or thread is externally terminated without its knowledge. The server bug may be a race between ps and process teardown. [653]

- There is a non-deterministic assertion failure in the server. Occasionally the server runs into an assertion failure when booting; apparently a port is being reused. [483]
- /dev/zero doesn't exist. The version in AD2 should be adopted. [277]
- Server priority depression incompatible with the new version of C-threads. In the new version, blocked C-threads no longer really block, they just depress their priority (with no timeout). Any server thread that depresses itself temporarily allow a different, unblocked server thread to run, will be put in the same priority run queue as all the blocked threads. Even if it is able to make forward progress almost immediately, it will not get any CPU time until all the blocked C-threads in front of it get their time first. [135]

3.4 Commands

- The df command uses signed ints for block counts. [2460]
- Ping 255.255.255.255 doesn’t work. [2361]
- The netstat command -I option returns “Interface not found” even though it can find the interface with different options (e.g. -i). [2326]
- The latest version of tcsh (6.06) sets the OSTYPE environment variable. However, the setting of OSTYPE interferes with the normal operation of the ODE tools. [2241]
- The -prune option to “find” doesn’t appear to work properly. If a directory is given as the argument to the -prune switch, find will descend into that directory. [2183]
- Running “ical” fails with: X Error of failed request: BadAlloc (insufficient resources for operation). Note that the ical window shows up and immediately vanishes. Failure occurs with X server in 16 bits per pixel mode. [2147]
- After a workstation has been up quite a while (a week or more), suddenly the X server refuses to create any new windows, claiming that the client is not authorized to connect to the server. Logging out and back in (i.e. restarting X) cures the problem. [2146]
- The newaliases command prints the error message “cannot alias non-local names”, while trying to alias local names. For example, the line “postmaster:root” in the aliases file generates this error. [2134]

- The X server crashes with an “Out of memory” error message. Problem can be worked around by adding the switch “-ld 32768” when starting the X server, This will increase X server’s data limit to 32M. [2082]
- If mwm is started without an .mwmrc, the system will dump core. The straightforward workaround is to insure you have a .mwmrc file, even if it is empty. [2073]
- Nslookup works fine normally, but when you ‘set type=any’ or fire it up as ‘nslookup -type=any’ then it fails with a record size incorrect error. [1904]
- Who command prints “tty??” instead of correct tty number. [1849]
- The gdb adapted to understand the new cthreads library should work better with the collocated servers. But in practice any significant use of it causes it to crash out with a memory allocation failure. [614]
- Using the -d option to pinfo (on an ordinary user-mode process) does not work. This is the option that uses *mach_port_space_info()* to query IPC space usage data. [103]

3.5 Libraries

- The cthreads implementation of red zones for cthread stacks wires the red zone memory permanently. This only occurs for cthreads stacks within colocated servers. It was observed that 14 pages get wired in the red zone for each in kernel stack. [2366]
- In the cthreads library the red_zone flag was forced to TRUE, even when the cthread_red_zone_size was not allocated, this leads to a vm_protect failure in stack_alloc. [2188]
- The rthreads library rthread_spawn primitive does not check for error return, and passes a NULL thread id to kernel (to get a ‘wired kernel thread’). [1977]

3.6 Unidentified

- On a two-node cluster, with a STD+WS+NORMA kernel from either NMK17.3 or cnmk_shared, you can reliably cause a multi-user boot of the AD1 server to fail by starting the boot process on both clusters almost simultaneously (node 0 a bit before node 1). So long as you get node 1 to complete registration for process creation before network configuration starts, the boot on node 0 will hang after outputting the hostname of node 0. If you bring up node 0 to multi-user and then add node 1 to the cluster, the failure does not occur. [564]

- When running two servers kernel-loaded, using the -2 option when starting the second server doesn't work. The second server crashes during the printing of the “Root” message. [105]