Mach: A New Kernel Foundation For UNIX Development

Sarthak Budhiraja, Rochan Mehrotra
Information Technology , Dronacharya College Of Engineering, Haryana,India

Abstract- Mach is a multiprocessor operating system kernel and environment under development at Carnegie Mellon University. Mach provides a new foundation for UNIX development that spans networks of uniprocessors and multiprocessors. This paper describes Mach and the motivations that led to its design. Also described are some of the details of its implementation and current status.

I. INTRODUCTION

Mach is a multiprocessor operating system kernel currently under development at Carnegie-Mellon University. In addition to binary compatibility with Berkeley's current UNIX 4.3BSD release, Mach provides a number of new facilities not available in 4.3:

• Support for multiprocessors including:
– provision for both tightly-coupled and loosely-coupled general purpose multiprocessors and
– separation of the process abstraction into tasks and threads, with the ability to execute multiple threads within a task simultaneously.
• A new virtual memory design which provides:
– large, sparse virtual address spaces,
– copy-on-write virtual copy operations,
– copy-on-write and read-write memory sharing between tasks,
– memory mapped files and
– user-provided backing store objects and pagers.
• A capability-based interprocess communication facility:
– transparently extensible across network boundaries with preservation of capability protection and
– integrated with the virtual memory system and capable of transferring large amounts of data up to the size of an address space via
copy-on-write techniques.
• A number of basic system support facilities, including:
– an internal adb-like kernel debugger,
– support for transparent remote file access between autonomous systems,
– language support for remote-procedure call style interfaces between tasks written in C, Pascal, and Common Lisp.

The basic Mach abstractions are intended not simply as extensions to the normal UNIX facilities but as a new foundation upon which UNIX facilities can be built and future development of UNIX-like systems for new architectures can continue. The computing environment for which Mach is targeted spans a wide class of systems, providing basic support for large, general purpose multiprocessors, smaller multiprocessor networks and individual workstations (see figure 1. As of April 1986, all Mach facilities, with the exception of threads, are operational and in production use on uniprocessors and multiprocessors by both individuals and research projects at CMU. In this paper we describe the Mach design, some details of its implementation and its current status.

II. DESIGN: AN EXTENSIBLE KERNEL

Early in its development, UNIX supported the notion of objects represented as file descriptors with a small set of basic operations on those objects (e.g., read, write and seek) [9]. With pipes serving as a program composition tool, UNIX offered the advantages of simple implementation and extensibility to a variety of problems. Under the weight of changing needs and technology, UNIX has been modified to provide a staggering number of different mechanisms for managing objects and resources. In addition to pipes, UNIX versions now support facilities such as System V streams, 4.2 BSD sockets, pty's, various forms of semaphores, shared memory and a mind-boggling array of ioctl operations on special files and devices. The result has been scores of additional system calls and options

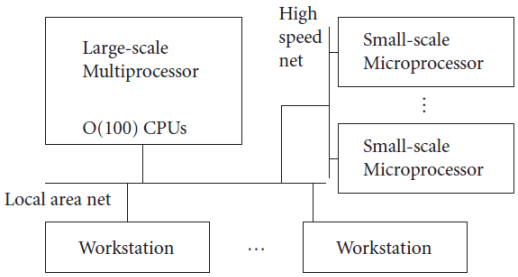

Figure 1: The Mach computing environment

with less than uniform access to different resources within a single UNIX system and within a network of UNIX machines.

As the complexity of distributed environments and multiprocessor architectures increases, it becomes increasingly important to return to the original UNIX model of consistent interfaces to system facilities. Moreover, there is a clear need to allow the underlying system to be transparently extended to allow user-state processes to provide services which in the past could only be fully integrated into UNIX by adding code to the operating system kernel.

The Mach kernel supports four basic abstractions:

1. A task is an execution environment in which threads may run. It is the basic unit of resource allocation. A task includes a paged virtual address space and protected access to system resources (such as processors, port
capabilities and virtual memory). The UNIX notion of a process is , in Mach, represented by a task with a single thread of control.

2. A thread is the basic unit of CPU utilization. It is roughly equivalent to an independent program counter operating within a task. All threads within a task share access to all task resources.

3. A port is a communication channel – logically a queue for messages protected by the kernel. Ports are the reference objects of the Mach design. They are used in much the same way that object references could be
used in an object oriented system. Send and Receive are the fundamental primitive operations on ports.

4. A message is a typed collection of data objects used in communication between threads. Messages may be of any size and may contain pointers and typed capabilities for ports. Operations on objects other than messages are performed by sending messages to ports which are used to represent them. The act of creating a task or thread, for example, returns access rights to the port which represents the new object and which can be used to manipulate it. The Mach kernel acts in that case as a server which implements task and thread objects. It receives incoming
messages on task and thread ports and performs the requested operation on the appropriate object. This allows a thread to suspend another thread by sending a suspend message to that thread's thread port even if the requesting thread is on another node in a network.

The design of Mach draws heavily on CMU's previous experience with the Accent [8] network operating system, extending that system's facilities into the multiprocessor domain:

• the underlying port mechanism for communication provides support for
object-style access to resources and capability based protection as well as
network transparency,

• all systems abstractions allow extensibility both to multiprocessors and to
networks of uniprocessor or multiprocessor nodes,

• support for parallelism (in the form of tasks with shared memory and
threads) allows for a wide range of tightly coupled and loosely coupled
multiprocessors and

• access to virtual memory is simple, integrated with message passing, and
introduces no arbitrary restrictions on allocation, deallocation and virtual copy operations and yet allows both copy-on-write and read-write sharing.

The Mach abstractions were chosen not only for their simplicity but also for performance reasons. A performance evaluation study done on Accent demonstrated the substantial performance benefits gained by integrating virtual memory
management and interprocess communication. Using similar virtual memory and IPC primitives, Accent was able to achieve performance comparable to UNIX systems on equivalent hardware [3].

III. TASKS AND THREADS

It has been clear for some time that the UNIX process abstraction is insufficient to meet the needs of modern applications. The definition of a UNIX process results in high overhead on the part of the operating system. Typical server applications, which use the fork operation to create a server for

each client, tend to use far more system resources than are required. In UNIX this includes process slots, file descriptor slots and page tables. To overcome this problem, many application programmers make use of coroutine packages to manage multiple contexts within a single process (see, for example, [2]).

Application parallelism in Mach can thus be achieved in any of three ways:

• through the creation of a single task with many threads of control executing in a shared address space, using shared memory for communication and synchronization,

• through the creation of many tasks related by task creation which share restricted regions of memory or

• through the creation of many tasks communicating via messages.

These alternatives reflect as well the different multiprocessor architectures to which Mach is targeted:

• uniform access, shared memory multiprocessors such as the VAX3 11/784, VAX 8300 and Encore MultiMax4,

• differential access shared memory machines such as the BBN Butterfly and IBM RP3,

• loosely-coupled networks of computers.

In fact, the Mach abstractions of task, thread and port correspond to the physical realization of many multiprocessors as nodes with shared memory, one or more processors and external communication ports.

IV. VIRTUAL MEMORY MANAGEMENT

The Mach virtual memory design allows tasks to:

• allocate regions of virtual memory,

• deallocate regions of virtual memory,

• set the protections on regions of virtual memory,

• specify the inheritance of regions of virtual memory. It allows for both copy-on-write and read/write sharing of memory between tasks. Copy-on-write virtual memory often is the result of form operations or large message transfers. Shared memory is created in a controlled fashion via an inheritance mechanism. Virtual memory related functions, such as pagein and pageout, may be performed by non-kernel tasks. Mach does not impose restrictions on what regions may be specified for these operations, except that they be aligned on system page boundaries (where the definition of the page size is a boot-time parameter of the system).

The way Mach implements the UNIX fork is an example of Mach's virtual memory operations. When a fork operation is invoked, a new (child) address map is created based on the old (parent) address map's inheritance values.

Inheritance may be specified as shared, copy or none, and may be specified on a per-page basis. Pages specified are shared, are shared for read and write access by both the parent and child address maps. Those pages specified as copy are effectively copied in the child map, however; for efficiency, copy-onwrite techniques are typically employed. An inheritance specification of none signifies that the page is not passed to the child at all. In this case, the child's corresponding address is left unallocated. By default, newly allocated memory is inherited copy-on-write.

Like inheritance, protection may be specified on a per-page basis. For each group of pages there exist two protection values: the current and maximum protection. The current protection controls actual hardware permissions. The maximum protection specifies the maximum value that the current protection may take. The maximum protection may never be raised, it may only be lowered.

If the maximum protection is lowered to a level below the current protection, the current protection is also lowered to that level. Either protection is a combination of read, write, and execute permissions. Enforcement of these permissions is dependent on hardware support (for example, many machines do not allow for explicit execute permissions, but those that do will be properly enforced).

V. VIRTUAL MEMORY IMPLEMENTATION

Given the wide range of virtual memory management built by hardware engineers, it was important to separate machine dependent and machine independent data structures and algorithms in the Mach virtual memory implementation.

In addition, the complexity of potential sharing relationships between tasks dictated clean separation between kernel data structures which manage physical resources and those which manage backing store objects.

The basic data structures used in the virtual memory implementation are:

address maps: doubly linked lists of map entries, each entry describing the properties of a region of

virtual memory. There is a single address map associated with each task.

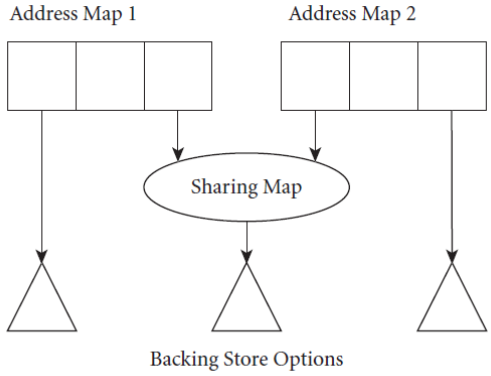

Figure 2: Task address maps

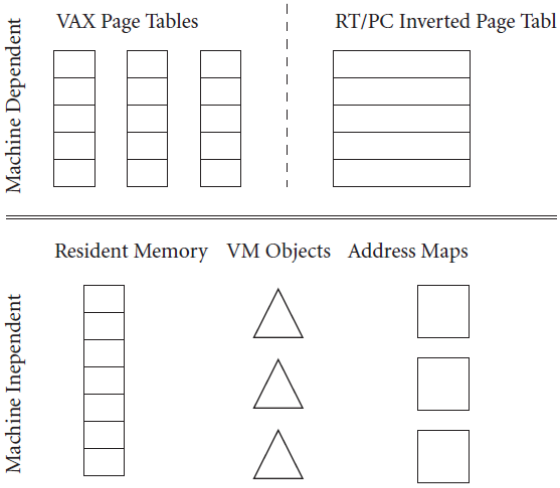

Figure 3: Task address maps

share maps: special address maps that describe regions of memory that are shared between tasks. A sharing map provides a level of indirection from address maps, allowing operations that affect shared memory to affect all maps without back pointers.

VM objects: units of backing storage. A VM object specifies resident pages as well as where to find non-resident pages. VM objects are pointed at by address maps. Shadow objects are used to hold pages that have been
copied after a copy-on-write fault.

page structures: specify the current attributes for physical pages in the system (e.g., mapped in what object, active/reclaimable/free).

The virtual memory implementation is split between machine independent and machine dependent sections. The machine independent portion of the implementation has full knowledge of all virtual memory related information.
The machine dependent portion, on the other hand, has a simple page validate/ invalidate/protect interface, and has no outside knowledge of other machine independent related data structures.
In addition to the normal demand paging of tasks, the Mach virtual memory
implementation allows portions of the kernel to be paged. In particular, address
map entries are pageable in the current implementation.

VI. INTERPROCESS COMMUNICATION

Interprocess communication in 4.3BSD can occur through a variety of mechanisms: pipes, pty's, signals, and sockets [7]. The primary mechanism for network communication, internet domain sockets, has the disadvantage of using global machine specific names (IP based addresses) with no location independence and no protection. Data is passed uninterpreted by the kernel as streams of bytes.
The Mach interprocess communication facility is defined in terms of ports and messages and provides both location independence, security and data type tagging.

The port is the basic transport abstraction provided by Mach. A port is a protected kernel object into which messages may be placed by tasks and from which messages may be removed. A port is logically a finite length queue of messages sent by a task. Ports may have any number of senders but only one receiver. Access to a port is granted by receiving a message containing a port capability (to either send or receive).
Messages may be sent and received either synchronously or asynchronously.
Currently, signals can be used to handle incoming messages outside the flow of
control of a normal UNIX style process. A task could create or assign separate
threads to handle asynchronous events.

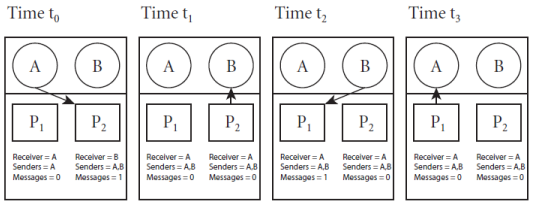

Figure 4: Typical message exchange

Figure 4 shows a typical message interaction. A task A sends a message to a port P2. Task A has send rights to P2 and receive rights to a port P1. At some later time, task B which has receive rights to port P2 receives that message which may in turn contain send rights to port P1 (for the purposes of sending a reply message back to task A). Task B then (optionally) replies by sending a message to P1.

Should port P2 have been full, task A would have had the option at the point of sending the message to: (1) be suspended until the port was no longer full, (2) have the message send operation return a port full error code, or (3) have the kernel accept the message for future transmission to port P2 with the proviso that no further message can be sent by that task to port P2 until the kernel sends a message to A telling it the current message has been posted.

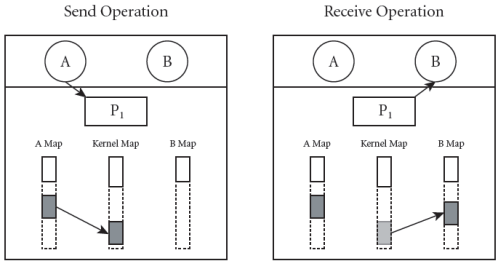

Figure 5: Memory mapping operations during message transfer

Figure 5 shows Task A sending a large (for example, 24 megabyte) message to a port P1. At the point the message is posted to P1, the part of A's address space containing the message is marked copy-on-write – meaning any page referenced for writing will be copied and the copy placed instead into A's virtual memory table. The copy-on-write data then resides in a temporary kernel address map
until task B receives the message. At that point the data is removed from the temporary address map. The operating system kernel determines where in the address space of B the newly received message data is placed, allowing the kernel to minimize memory mapping overhead. Any attempt by either A or B to change a page of this copy-on-write data results in a copy of that page being made and placed into that task's address space.

VII. SYSTEM SUPPORT FACILITIES

In addition to the basic system support facilities provided by 4.3, Mach provides a kernel debugger and a transparent remote file system.

7.1 Kernel Debugger

Kernel debugging has always been a tedious undertaking. UNIX systems traditionally have no support for kernel debugging, requiring kernel implementers to "debug with printfs" or other ad hoc methods. The Mach kernel has a builtin kernel debugger (kdb) based on adb7. All adb commands are implemented including support for breakpoints, single instruction step, stack tracing and symbol table translation.

In order to aid debugging, as well as study the performance of the kernel, the Mach debugger also supports functions not available in adb. For example: enhanced stack traces: stack traces may contain the values of local variables and registers for each stack frame. call/return trace support: single stepping may continue without intervention until the next call or return instruction.

instruction counting: the number of instructions executed between regions of code may be counted.

During the implementation of the system these features have proved invaluable in both debugging and performance tuning.

7.2 Transparent Remote File system

The remote file system available in Mach was originally available in 1982 as part of CMU's locally maintained version of 4.1 UNIX. At that time, it supported only a small set of the functions required of a file system: it could read and/or write publicly accessible files. Over the years, the remote filesystem has undergone a steady increase in functionality. Currently, all UNIX functions, such as remote current directories and execution of remote files, are supported. The remote filesystem is completely transparent to the user. Users may
effectively login to a remote filesystem connection to receive all of their normal privileges on the remote filesystem, or they may elect to not login, and receive only "anonymous" access to the remote filesystem. A small set of kernel hooks redirects remote file operations to remote servers transparently. Each machine wishing to allow

remote requests runs a usermode server process. The kernel sends requests corresponding to operations such as read, write, open and close. The client then performs the appropriate operation, and returns with a reply code and/or data. Data is not cached with one exception: remote execution of files causes a cached copy of the entire file to be read into an inode on a local disk. Subsequent executions of this file cause the kernel to check for a modification of the remote file; if no such modification
has been made, then the locally cached copy is executed.

VIII. IMPLEMENTATION: A NEW FOUNDATION FOR UNIX

The Mach kernel currently supplants most of the basic system interface functions of the UNIX 4.3BSD kernel: trap handling, scheduling, multiprocessor synchronization, virtual memory management and interprocess communication. 4.3BSD functions are provided by kernel-state threads which are scheduled by the Mach kernel and share communication queues with it.

The spectacular growth in size of the Berkeley UNIX kernel over the last few years has made it apparent that continued expansion of UNIX functionality threatens to undercut the advantages of simplicity and modifiability which mad UNIX an attractive operating system alternative for research and development.

Work is underway to remove non-Mach UNIX functionality from kernel-state and provide these services through user-state tasks. The goal of this effort is to "kernelize" UNIX is a substantially less complex and more easily modifiable basic operating system. This system would be better adapted to new uniprocessor and multiprocessor architectures as well as the demands of a large network
environment. The success of this transition will depend heavily on the fact that the basic Mach abstractions allow kernel facilities such as memory object management and interprocess communication to be transparently extended. Figure 6 shows the eventual relationship between the Mach kernel and UNIX.

IX. CURRENT STATUS: MACH-1

Mach is still under development and extensive performance comparisons with other systems have not yet been done. Although the system has yet to be tuned, current performance appears to be in line with 4.3BSD. Some early simplistic measures of virtual memory performance are encouraging. The MicroVAX II cost of touching newly allocated memory is less than 0.7 milliseconds per 1024 bytes of data (versus approximately 1.2 milliseconds for 4.3BSD). Operations typically expensive in UNIX, e.g. fork, are substantially faster with the new virtual memory support. Mach is currently in production use by CMU researchers on a number of projects including a multiprocessor speech recognition system
called Agora and a project to build parallel production systems.

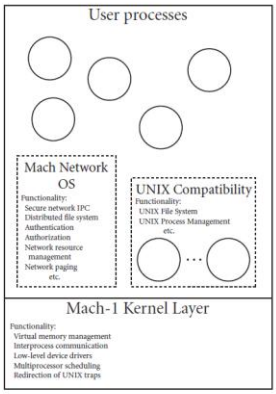

Figure 6: Mach with UNIX functionality in user-state tasks. As of April 1986 the box labeled "UNIX compatibility" still executes in kernel state and communicates with the Mach kernel layer through a shared communication queue.

As of April 1986, Mach runs on most VAX architecture machines: VAX 11/750, 11/780, 11/785, 8600, MicroVAX I, and MicroVAX II. In addition, Mach runs on four (11/780 or 11/785) processor VAX 11/784 with 8 MB of shared memory and the IBM RT/PC. The same binary kernel image runs on all VAX uniprocessors and multiprocessors. The same kernel source is used for both VAX and RT/PC systems. Work has begun on ports to the uniprocessor
SUN 3, multiprocessor Encore MultiMax and VAX 8300. Implementation of the Mach thread mechanism is expected by Summer 1986.

REFERENCES

[1] Wikipedia : Mach for UNIX

[2] D. R. Brownbridge, L.F. Marshall, and B. Randell. The newcastle connection, or UNIXes of the world unite! Software - Practice and Experience, 20, 1982.

[3] M. Satyanarayanan et al. The ITC distributed file ssystem: Principles and design. pages 35–50. ACM, December 1985.

[4] R. Fitzgerald and R. F. Rashid. The integration of virtual memory management and interprocess communication in accent. ACM Transactions on Computer Systems, 4(2), May 1986.
[5] A. K. Jones. The object model: A conceptual tool for structuring systems. Operating Systems: An Advanced Course, pages 7–16, 1978.
[6] M. B. Jones, R. F. Rashid, and M. Thompson. Matchmaker: An interprocess specification language. ACM, January 1985.