The Mach System


Presented by James Holladay
Outline

- How Mach started
- Goals of Mach
- Benefits
- Primitive Abstractions
- Memory and IPC
- C Threads Package
- CPU Scheduler
Outline (Continued)

- Memory
- Memory Managers
- Shared Memory
- Summary
How Mach Started

- Mach traces its ancestry to the Accent operating system developed at Carnegie Mellon University
  - communication system and philosophy are derived from Accent
- Unlike Accent, Mach is:
  - Able to execute UNIX Applications
  - Not tied to any one architecture
How Mach Started (Continued)

• Mach code was first developed inside the 4.2BSD kernel
  ◦ Mach components replaced BSD ones as they were completed

• Mach 3 moves BSD code outside the kernel
  ◦ Microkernel
  ◦ Allows replacement of BSD with another OS
    • Or, the simultaneous execution of multiple operating-system interfaces on top of the microkernel
Goals of Mach

- **Compatibility with UNIX**
  - Mach is compatible with UNIX 4.3BSD

- **Support diverse architectures**
  - Varying number of processors (to thousands)
  - Varying degrees of shared memory access

- **Simplified kernel structure**
  - Small number of abstractions
  - Minimize code within the kernel
  - Make the code powerful enough that all other features can be implemented at user level
Mach’s Benefits

- Simple kernel structure and abstractions
  - General enough to allow other operating systems to be implemented on top of Mach
  - Avoids having too many competing ways to perform the same task
- Example of this simplification:
  - All requests to the kernel, and all data movement among processes, are handled through one communication mechanism
    - Mach is able to provide system wide protection by protecting the communications mechanism
    - Optimizing this communications path can increase performance, and is simpler than optimizing several paths
Mach’s Primitive Abstractions

- Task (Execution Environment)
- Thread (Unit of Execution)
- Port (Object Reference Mechanism)
- Port Set
- Message (Thread Communication)
- Memory Object (Source of Memory)
Mach’s Primitive Abstractions

Figure B.2  Mach’s basic abstractions.
Mach Primitives - Tasks

- A **task** is an *execution environment* that provides the basic unit of resource allocation
  - Virtual address space
  - Protected access to system resources via ports
  - A task can contain 1 or more threads
  - States: **Running, Suspended** (explained on next slide)

- An operation on a task affects all threads in a task
  - Suspending a task suspends all the threads in it
  - Task and thread suspensions are separate, independent mechanisms
    - Resuming a thread in a suspended task does not resume the task

- A **task** can be thought of as a traditional process that does not have an instruction pointer or a register set.
  (but the task does nothing without 1+ threads)
Mach Primitives - Threads

- A **thread** is the **basic unit of execution**
  - Must run in a task (which provides the address space)
  - All threads within a task share the tasks’ resources
    - Ports
    - Memory

- **States:**
  - **Running:**
    - Thread is executing
    - Waiting to be given a CPU
    - A thread is considered to be running even if it is blocked within the kernel (a page fault, etc.)
  - **Suspended:**
    - Thread is not executing
    - Not waiting to be given a CPU
    - The thread can resume only if it is returned to the “running” state
Mach Primitives – Ports, Port Sets

- A **port** is the basic **object reference mechanism** in Mach
  - It is a kernel-protected communication channel
    - Communication: sending messages to ports
    - A message is queued at the destination port if no thread can receive it
    - Ports are protected by **port rights** *(required to send message)*
- The programmer invokes an operation on an object by **sending** a message to a port associated with that object
- The object being represented by a port **receives** the messages
- A **port set** is a group of ports sharing a **common message queue**
  - A thread can receive messages for a port set, and thus **service multiple ports**
    - Each received message identifies the individual port (within the set) that it was received from; the receiver can use this to identify the object referred to by the message
Ports (Continued)

- A port is a protected, bounded queue within the kernel. If full, a sender may abort the send, wait for a slot to become available, or have the kernel deliver the message for it.

- Allocate a new port for a task
  - Task given all access rights to the port
  - Port name is returned

- Deallocate rights to a port
  - If task is destroyed that is receiving
    - Destroy port, all other sending to that port (potentially) notified

- Ports created by the kernel for a new task:
  - Task_self – handle’s the task’s kernel calls.
  - Task_notify – receives notification messages for the task (eg. If a port is closed)

- *Mach ports can be transferred only in messages*
Port Security

- Mach ensures security by requiring that message senders and receivers have *rights*:
  - A port name
  - A capability (send or receive) on that port
  - Only one task with receive rights to any given port
    - Allows IPC to be used for synchronization (passing a resource with messages)
  - Many tasks may have send rights

- When an object is created:
  - New port to represent the object
  - Creator obtains the access rights
    - Rights can be given out by the creator, and are passed in messages
      - If the holder of a receive right sends that right in a message, the receiver of the message gains the right and the sender loses it.

- A task may allocate ports:
  - To allow access to any objects it owns, or for communication

- The destruction of either a port or the holder of the receive right causes:
  - Revocation of all rights to that port
  - Tasks holding send rights can be notified
Mach Primitives - Messages

- A **message** is the basic method of communication between threads in Mach
  - *Typed* data object(s)
    - Actual data
    - A pointer to out-of-line data
    - Port rights
      - Passing port rights in messages is the only way to move them among tasks. (Passing a port right in shared memory does not work, because the Mach kernel will not permit the new task to use a right obtained in this manner.)
Mach Primitives - Messages

Figure B.5  Mach messages.
Messages Between Hosts

- The kernel uses the NetMsgServer when a message needs to be sent to a port that is not on the kernel’s computer:
  1. Mach’s kernel IPC to the local NetMsgServer
  2. Local NetMsgServer to remote NetMsgServer by an appropriate protocol
  3. Remote NetMsgServer uses that kernel’s IPC to send the message to the correct destination task

- As a security precaution, a port value provided in an add request must match that in the remove request for a thread to ask for a port name to be removed from the database
Message Passing Diagram

Figure B.7  Network IPC forwarding by NetMsgServer.
Mach Primitives – Memory Objects

- A **memory object** is a source of memory; tasks may access it by mapping portions (or the entire object) into their address spaces
  - May be managed by a user-mode *external memory manager*
    - Example: a file managed by a file server
  - A memory object can be any object for which memory-mapped access makes sense
Mach Primitives – Memory Objects

- A secondary-storage object is usually mapped into the virtual address space of a task
- Thread access -> fault: kernel sends a memory object data request message to the memory object’s port
  - The thread is placed in wait state until the memory manager either
    - Returns the page in a memory object data provided call
    - Returns an appropriate error to the kernel
  - …Meaning memory objects can be created and serviced by non-kernel tasks

- The end result is that, in the traditional sense, memory can be paged by user-written memory managers. When the object is destroyed, it is up to the memory manager to write back any changed pages to secondary storage.
Memory and IPC Integration

- IPC in Memory
  - Object is represented by a port (or ports), and IPC messages are sent to this port to request operations

- Memory in IPC
  - Where possible, Mach passes messages by moving pointers to shared memory objects, rather than by copying the object itself
C Threads Package

- Create a new thread within a task
  - Runs concurrently with the calling thread; calling thread receives a thread ID
- Destroy the calling thread, and return a value to the creating thread
- Wait for a specific thread to terminate before allowing the calling thread to continue
  - This is a synchronization tool
- Yield use of a processor; increases efficiency
- Mutual Exclusion in C Threads:
  - `mutex alloc`
  - `mutex free`
  - `mutex lock`
  - `mutex unlock`
C Threads Package (Continued)

- General synchronization without busy waiting can be achieved through the use of *condition variables*, which can be used to implement a *monitor*.

- **Condition_alloc**
- **Condition_free**
- **Condition_wait** unlocks the associated mutex variable
  - Blocks the thread until a *condition signal* is executed on the condition variable
    - Mutex variable is then locked, and the thread continues

- **condition signal**
  - Does not guarantee that the condition still holds when the unblocked thread finally returns from its *condition wait* call, so the awakened thread must loop, executing the *condition wait* routine until it is unblocked and the condition holds.
Mach’s CPU Scheduler

- The CPU scheduler for a thread-based multiprocessor OS is more complex than a process-based one
  - Generally more threads in a multithreaded system than there are processes in a multitasking system
  - Tracking multiple processors is difficult

- Mach uses a simple policy to keep the scheduler manageable:
  - Only threads are scheduled, so no knowledge of tasks is needed in the scheduler
  - All threads compete equally for resources, including time

- Each thread has a priority number ranging from 0 through 127, which is based on its CPU usage
  - A thread that recently used the CPU for a long time has the lowest priority
  - Priority places the thread in one of 32 global run queues
    - These queues are searched in priority order for waiting threads when a processor becomes idle
  - Mach also keeps per-processor, or local, run queues
    - Used for threads that are bound to an individual processor
Each processor consults the run queues to select the next thread to run
  ◦ Threads in the local run queue have priority over those in the global queues
Run queues are locked when they are modified to avoid corruption
Mach maintains a list of idle processors to serve the global run queue

Mach varies the size of the interrupt timing inversely with the total number of threads in the system
  ◦ More threads? Shorter time between interrupts
No need to interrupt if fewer threads than processors…

Disruptions come in two varieties:
  ◦ Internally generated exceptions
    • Asynchronously generated disruptions of a thread or task
  ◦ External interrupts
    • Unusual conditions during a thread’s execution
Memory

- The virtual address space of a task is generally sparse
  - Holes of unallocated space
- Mach makes no attempt to compress the address space
- A task may fail (crash) if it has no room for a requested region in its address space
  - Memory is cheap
- Page-table space is used for only currently allocated regions
- Page fault:
  - Kernel checks to see whether the page is in a valid region, rather than simply indexing into the page table and checking the entry
    - Reduced memory-storage requirements
    - Simpler address-space maintenance
Memory Managers

Responsible for the consistency of the contents of a memory object mapped by tasks on different machines

- **Useful for:**
  - Maintaining consistency of secondary storage for threads on multiple processors
  - Controlling the order of operations on secondary storage, to enforce consistency constraints demanded by database management systems

- **Insufficient in cases where:**
  - A task allocating a new region of virtual memory might not have a memory manager assigned to that region, since it does not represent a secondary-storage object
Shared Memory

- All threads in a task share that task’s memory, so no formal shared-memory facility is needed within a task
- Parent task can declare memory regions that can be inherited by its children for reading and writing
  - All changes are made to the same copy
  - Threads handle synchronization
Summary

- Mach (at least in version 3) has minimized the code left in the kernel, allowing operating system(s) to exist on top of it.
- Mach uses few, but useful abstractions, from which all other functionality can be built.
- Mach can be used on multiple computer architectures, and for many CPUs.
Summary

- Mach allows that OS on top of it to be made up of user-level tasks
  - Message passing is very useful