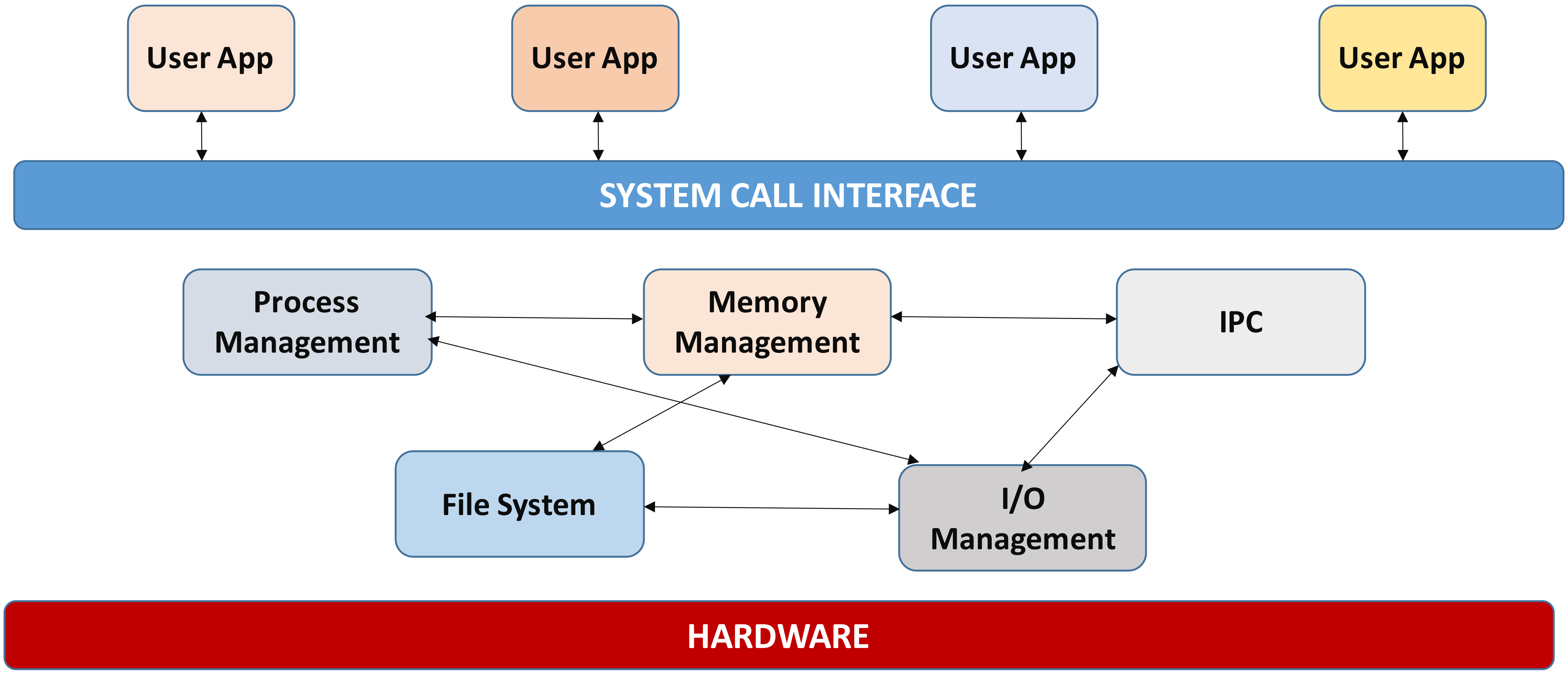
XNU Overview

Kenel Structure and Task Scheduling

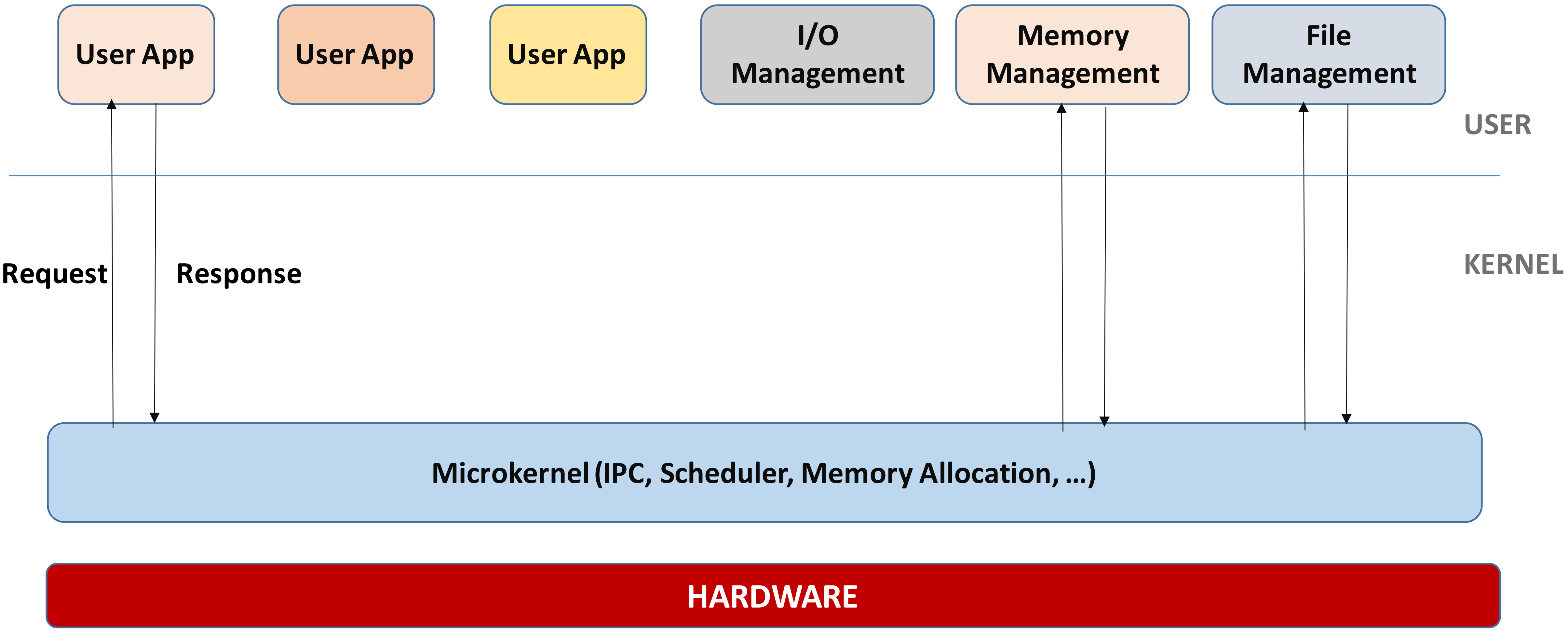
Department of Computer Science and Engineering

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

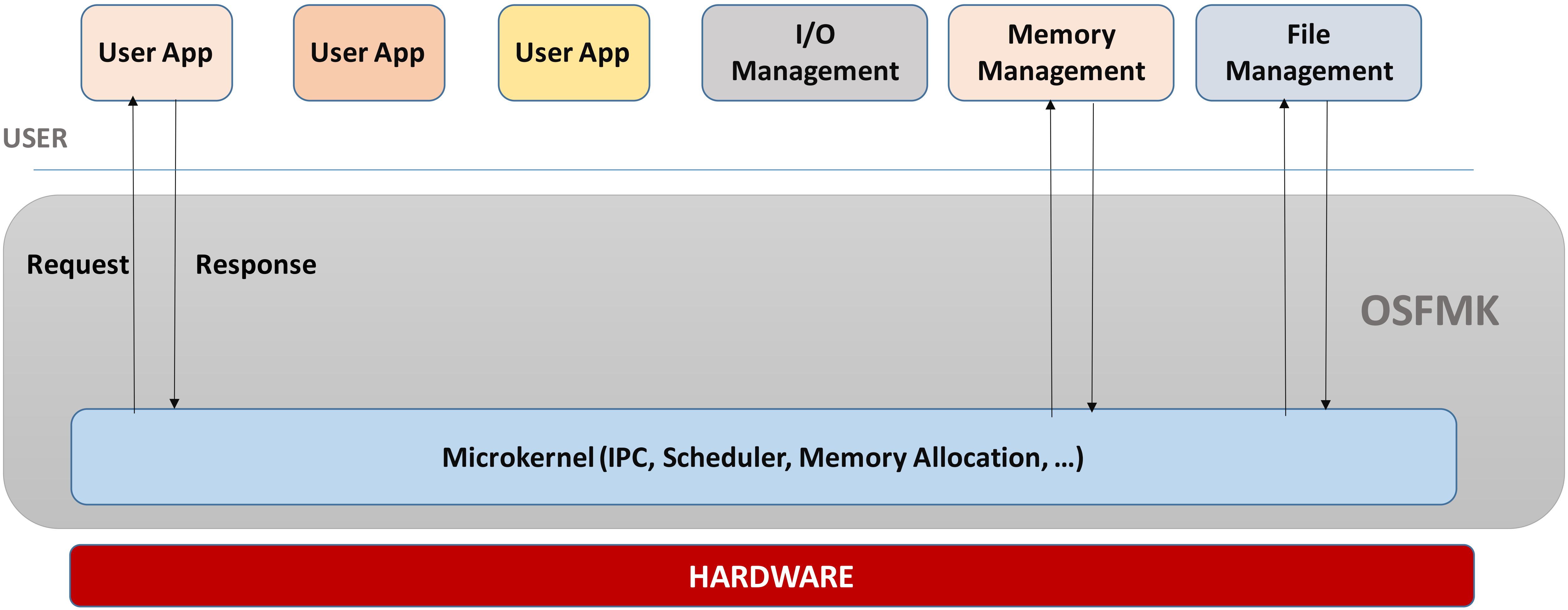
Monolithic Kernel



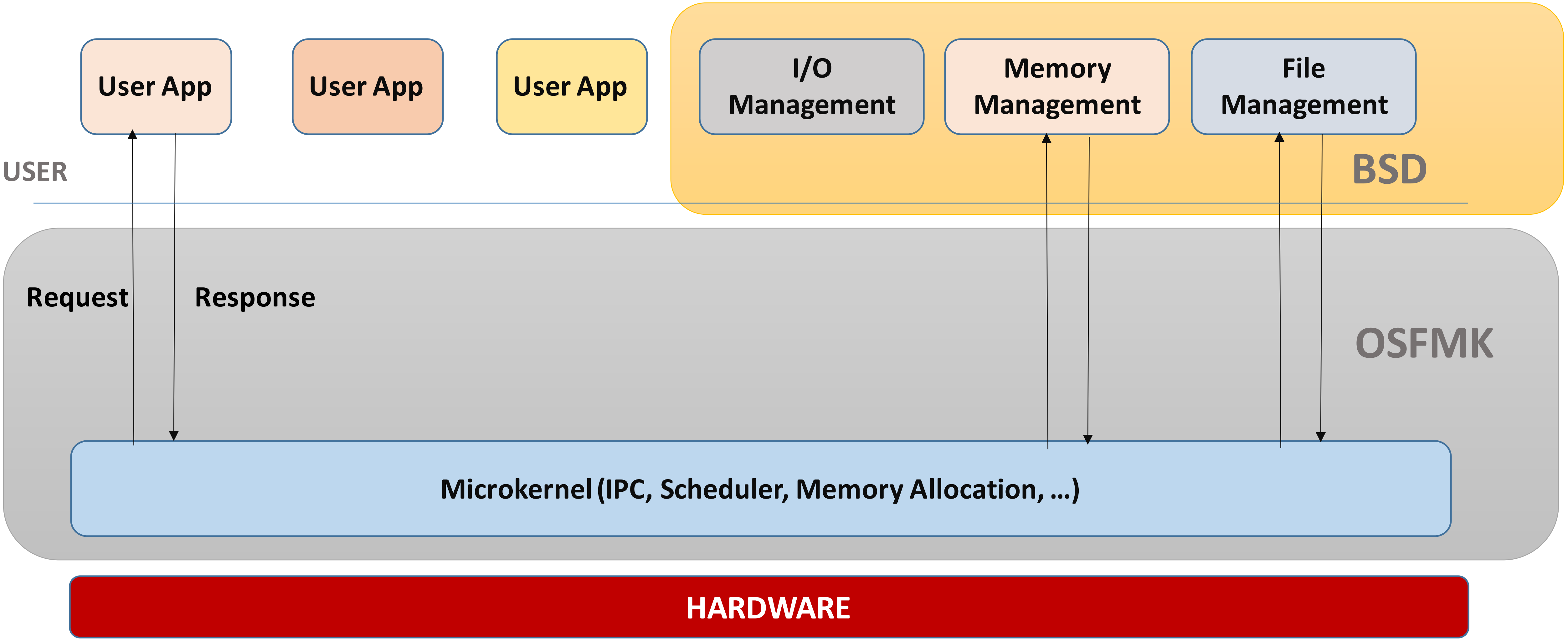
Microkernel



XNU Abstraction



XNU Abstraction

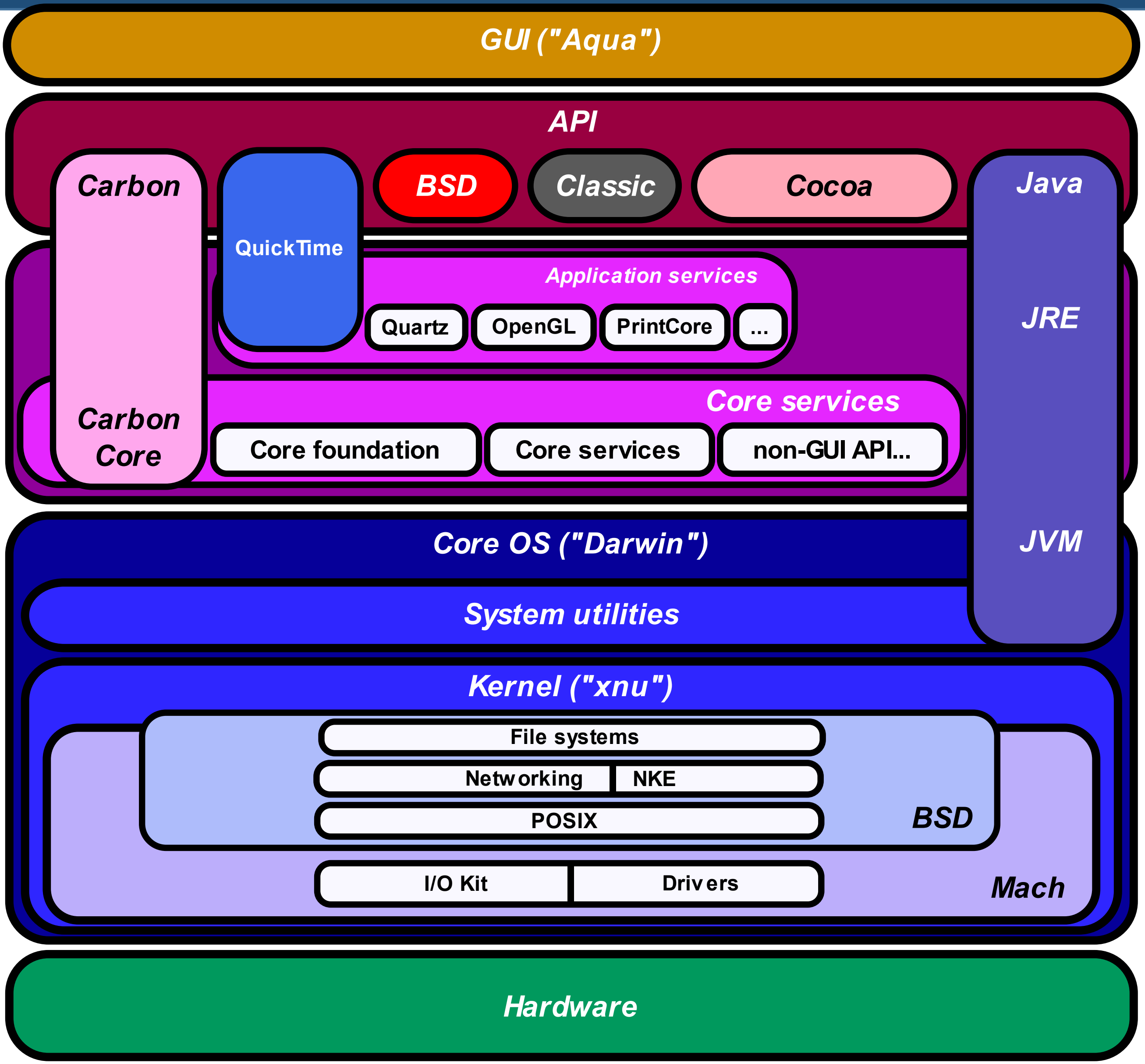


XNU: A Bit of Background

- Originally developed by NeXT for NeXTSTEP Operating System
  - MACH 2.5 (developed by CMU) + 4.3 BSD + Objective C for writing drivers (Driver kit)
- After Apple acquired NeXT,
  - Updated the MACH code with OSFMK 7.3 (ported from MACH 4 developed by University of Utah and MACH 3.0 forked from MACH 2.5) + FreeBSD + Embedded C++ for drivers (IO Kit)
- XNU Source Tree: https://github.com/apple-oss-distributions/xnu
- Darwin OS: The core of macOS (previously OS X), released in 2000, uses XNU kernel plus the application interfaces

The Darwin OS -- the Core of Apple OS

audioOS - Homepod
macOS - iMac
bridgeOS - Apple HW
iOS - iPhone/iPad
tvOS - Apple TV
watchOS - Apple Watch



MACH Primitives

- It has a highly minimalist concept:
  - A thin, minimal core, supporting an object-oriented model wherein individual, well-defined components (in effect, subsystems) communicate with one another by means of messages.
- Unlike other OSes that present a complete model on top of which user mode processes may be implemented, Mach provides a bare-bones model, on top of which the OS itself may be implemented.
- Mach is, essentially, a kernel-within-a kernel.
  - While Mach calls are visible from user mode, they implement a deep core, on top of which a larger kernel may be implemented.
- XNU is one specific implementation of UNIX (specifically, BSD) over Mach

XNU Source: https://opensource.apple.com/source/xnu/

Everything is Object

- In Mach, everything is implemented as its own object.
- Processes (which Mach calls tasks), threads, and virtual memory are objects, each with its own properties.
- Mach implements object-to-object communication by means of message passing.
- Unlike other architecture, Mach objects cannot directly invoke or call on one another, rather they are required to pass messages.

Message Passing

- The source object sends a message, which is queued by the target object until it can be processed and handled.
- Similarly, the message processing may produce a reply, which is sent back by means of a separate message
- Messages are delivered reliably in a FIFO manner.
- The content of the message is entirely up to the sender and the receiver to negotiate.

Mach Kernel Abstractions

- All functionalities are moved out of the kernel into user mode, leaving the kernel with the bare minima.
- Mach provides a small set of abstractions that have been designed to be both simple and powerful.

Mach Kernel Abstractions

- These are the main kernel abstractions:

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups

Indian Institue of Technology Kharagpur

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.
  - **Address space -** In conjunction with memory managers, Mach implements the notion of a sparse virtual address space and shared memory.

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.
  - **Address space -** In conjunction with memory managers, Mach implements the notion of a sparse virtual address space and shared memory.
  - **Memory objects** - The internal units of memory management. Memory objects include named entries and regions; they are representations of potentially persistent data that may be mapped into address spaces.

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.
  - **Address space -** In conjunction with memory managers, Mach implements the notion of a sparse virtual address space and shared memory.
  - **Memory objects** - The internal units of memory management. Memory objects include named entries and regions; they are representations of potentially persistent data that may be mapped into address spaces.
  - **Ports -** Secure, simplex communication channels, accessible only via send and receive capabilities (known as port rights).

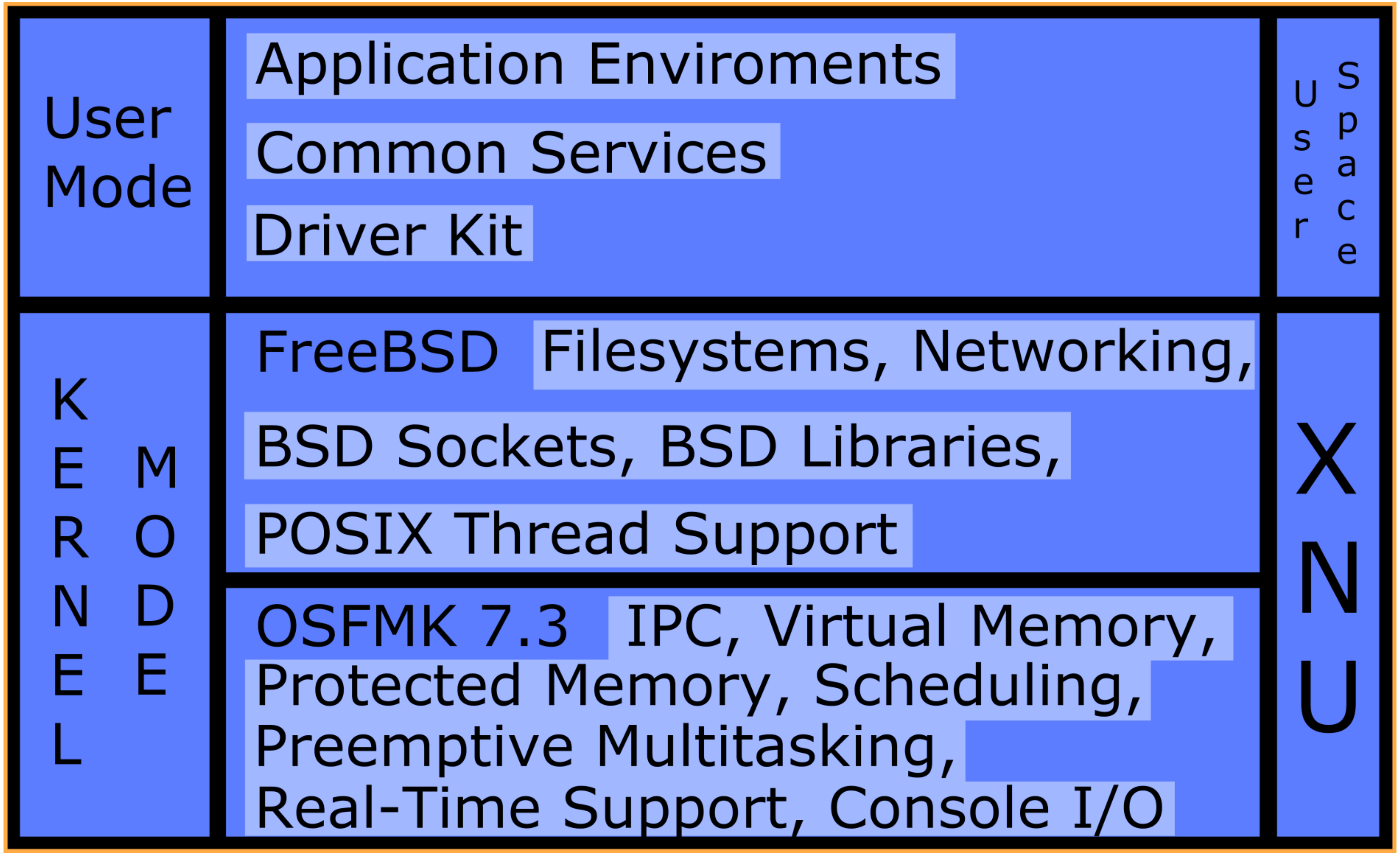
Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.
  - **Address space -** In conjunction with memory managers, Mach implements the notion of a sparse virtual address space and shared memory.
  - **Memory objects** - The internal units of memory management. Memory objects include named entries and regions; they are representations of potentially persistent data that may be mapped into address spaces.
  - **Ports -** Secure, simplex communication channels, accessible only via send and receive capabilities (known as port rights).
  - **IPC -** Message queues, remote procedure calls, notifications, semaphores, and lock sets.

Mach Kernel Abstractions

- These are the main kernel abstractions:
  - **Tasks -** The units of resource ownership; each task consists of a virtual address space, a port right namespace, and one or more threads(Similar to a process.). Allocation of resources to individual threads or groups
  - **Threads** - The units of CPU execution within a task.
  - **Address space -** In conjunction with memory managers, Mach implements the notion of a sparse virtual address space and shared memory.
  - **Memory objects** - The internal units of memory management. Memory objects include named entries and regions; they are representations of potentially persistent data that may be mapped into address spaces.
  - **Ports -** Secure, simplex communication channels, accessible only via send and receive capabilities (known as port rights).
  - **IPC -** Message queues, remote procedure calls, notifications, semaphores, and lock sets.
  - **Time -** Clocks, timers, and waiting.

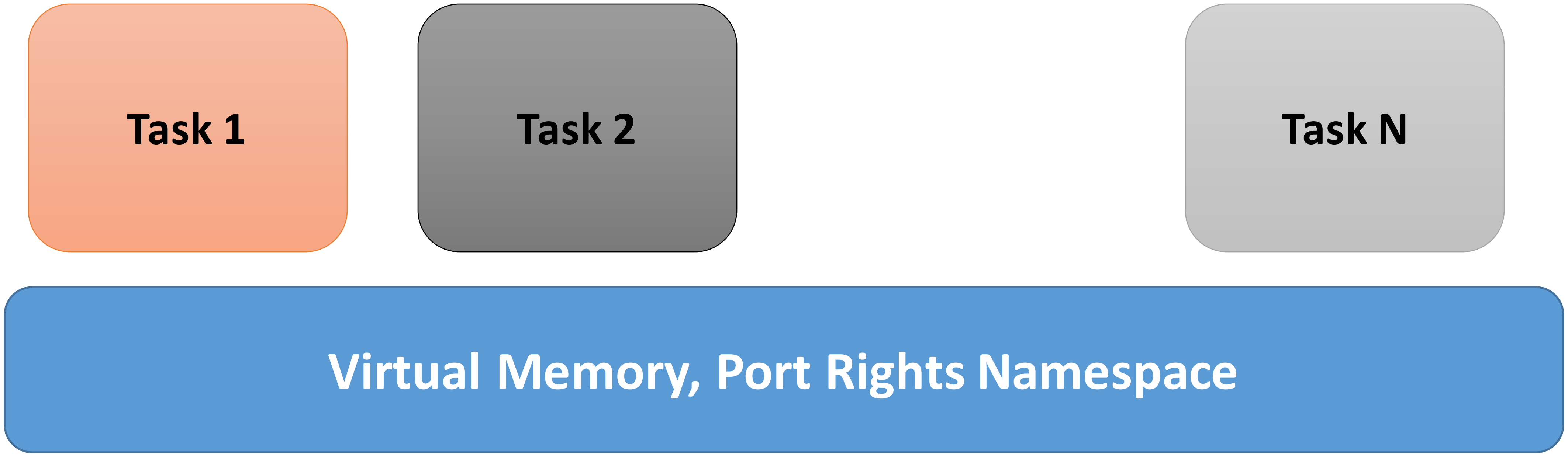
XNU Abstraction



Tasks and Threads

Virtual Memory, Port Rights Namespace

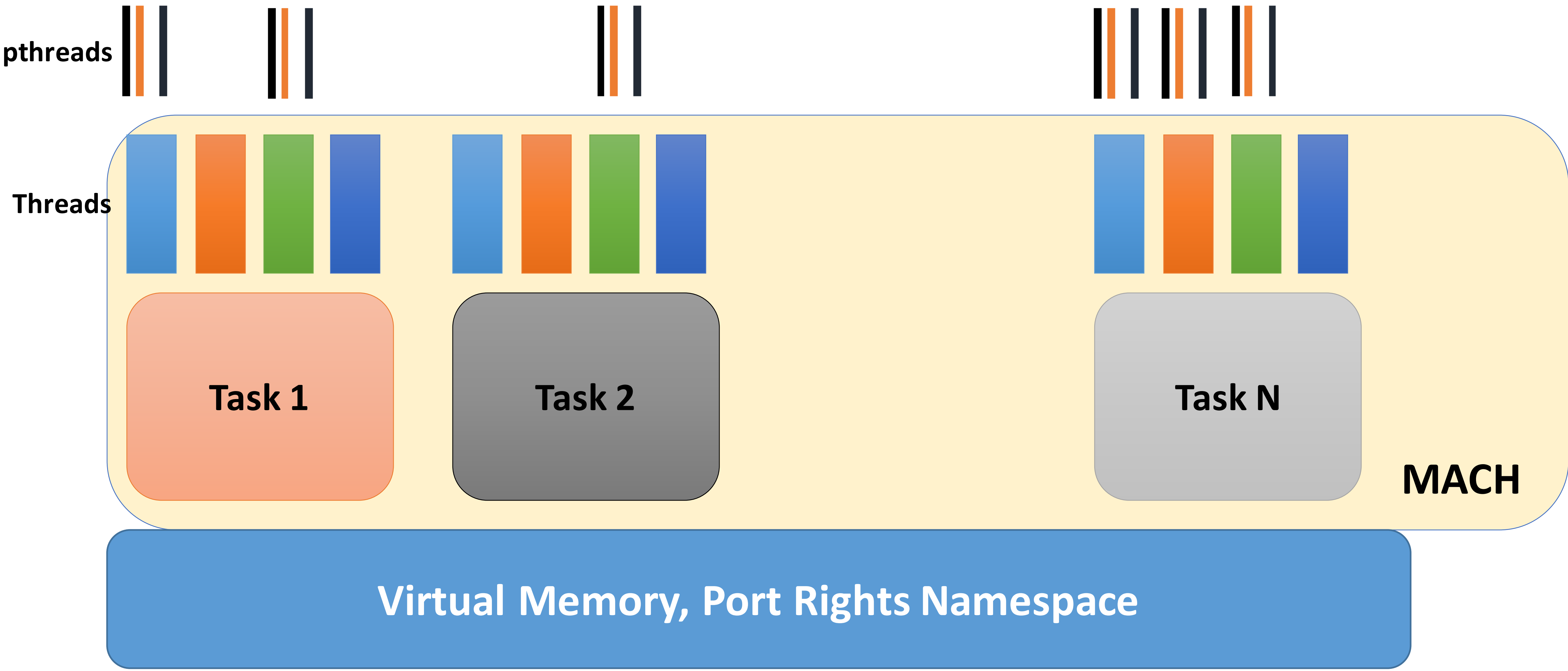
Tasks and Threads



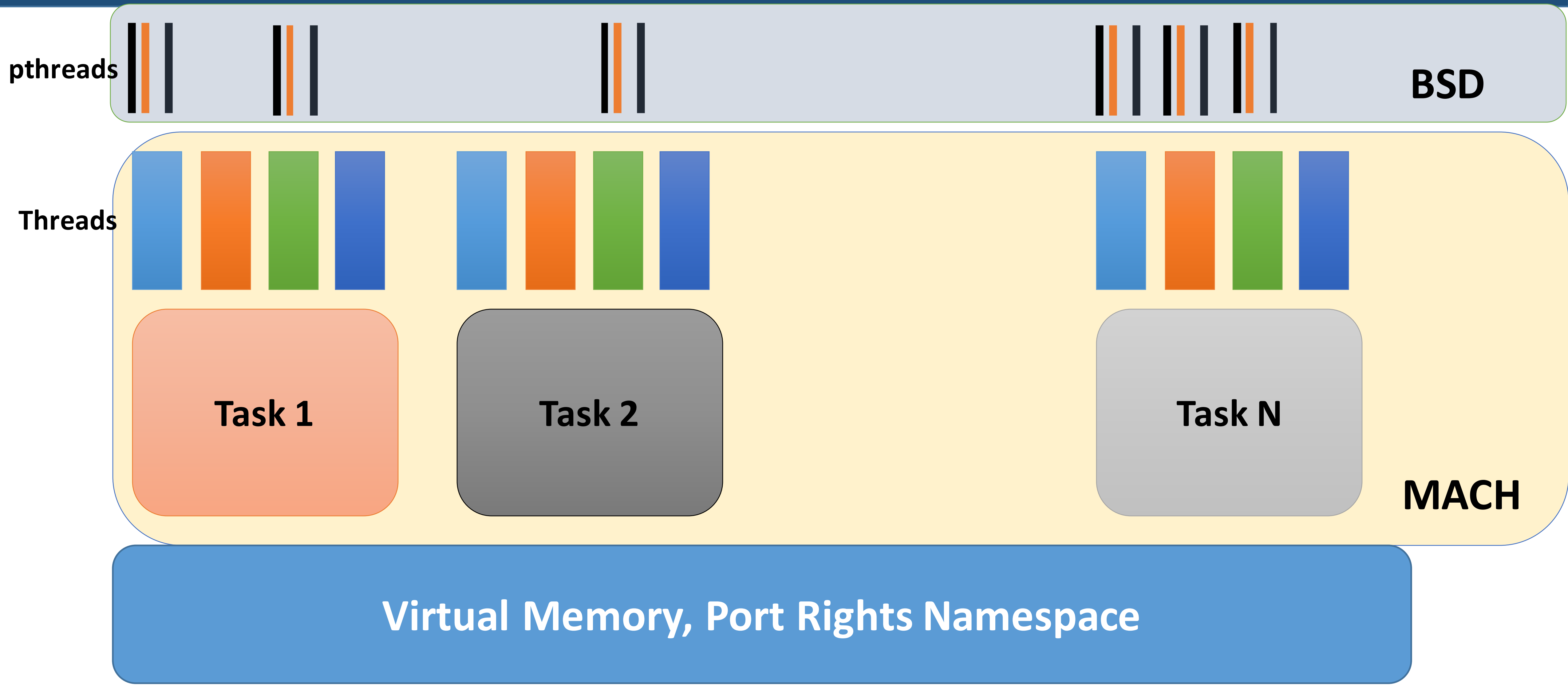
Tasks and Threads

Tasks and Threads

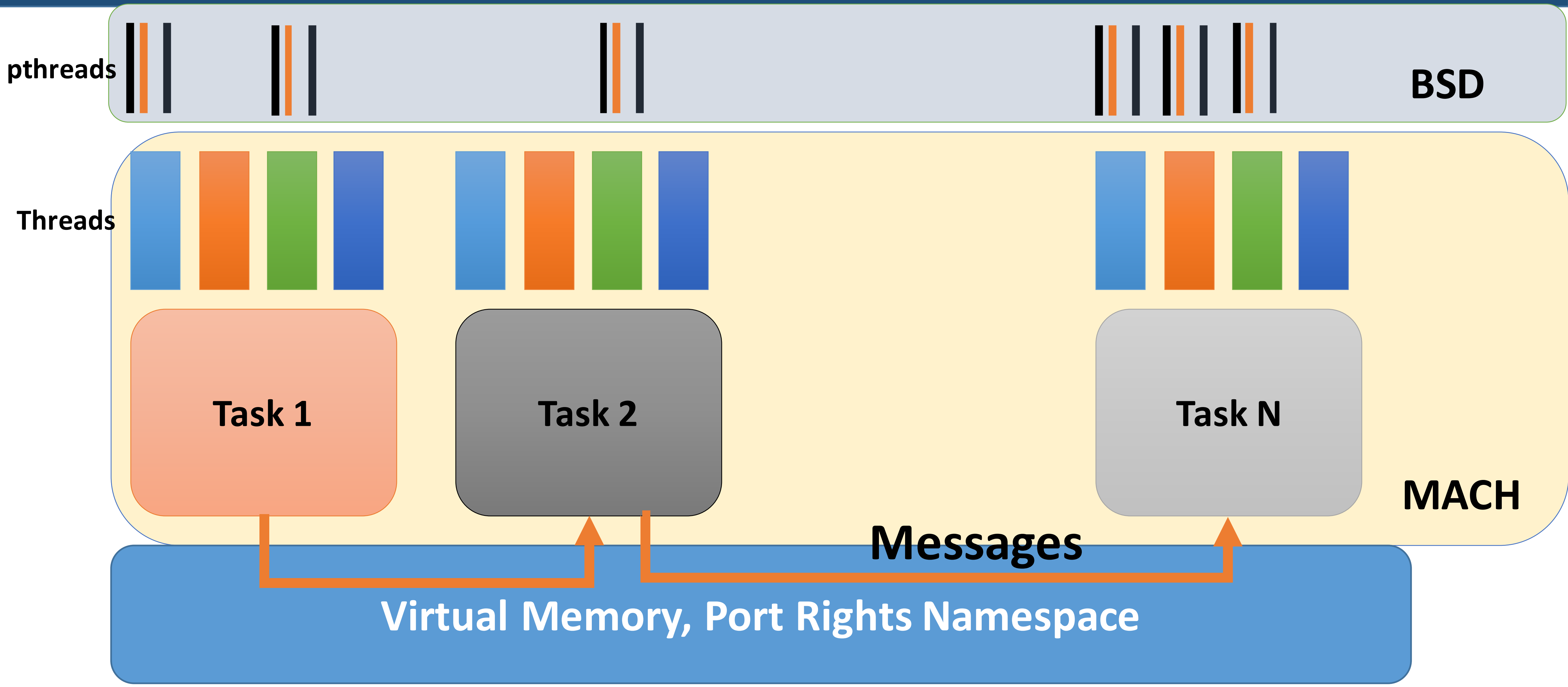
Tasks and Threads



Tasks and Threads



Messages



pthreads
BSD
Threads
Task 1
Task 2
Task N
MACH
Messages
Virtual Memory, Port Rights Namespace

SCHEDULING PRIMITIVES

- It describes tasks and threads, and the application programming interfaces (APIs) they offer.
- The kernel sees threads, not processes.
- It uses the concepts of the more lightweight tasks rather than processes.
- Classic UNIX uses a top-down approach, in which the basic object is a process that is further divided into one or more threads.
- Mach, on the other hand, uses a bottom-up approach in which the fundamental unit is a thread, and one or more threads are contained in a task.

Threads

- A thread defines the atomic unit of execution in Mach.
- It represents the underlying machine register state and various scheduling statistics.
- Defined in kern/thread.h (https://github.com/apple/darwin-xnu/blob/main/osfmk/kern/thread.h)
  - A thread is designed to provide the maximum information required for scheduling, while maintaining the lowest overhead possible.
- A thread contains no actual resource references.
- Mach defines the task as a thread container, and it is the task level in which resources are handled.
- A thread has access (via ports) to only the resources and memory allocated in its containing task.

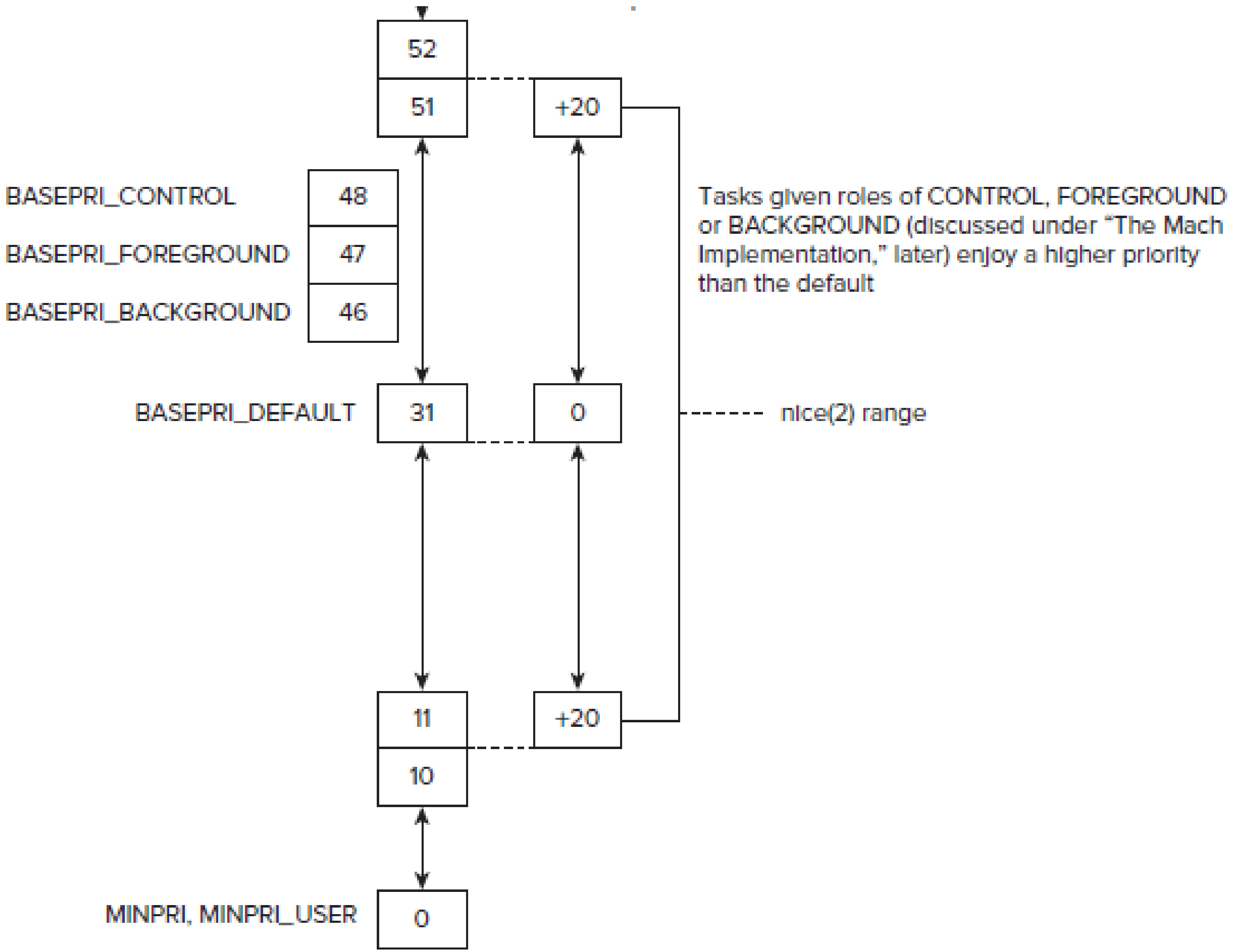
Tasks

- A *task* serves as a container object, under which the virtual memory space and resources are managed.
- These resources are devices and other handles.
- The resources are further abstracted by ports.
- Sharing resources thus becomes a matter of providing access to their corresponding ports.
- The task is a relatively lightweight structure (at least, compared to the threads), defined in osfmk/kern/task.h (https://github.com/apple/darwin-xnu/blob/main/osfmk/kern/task.h)

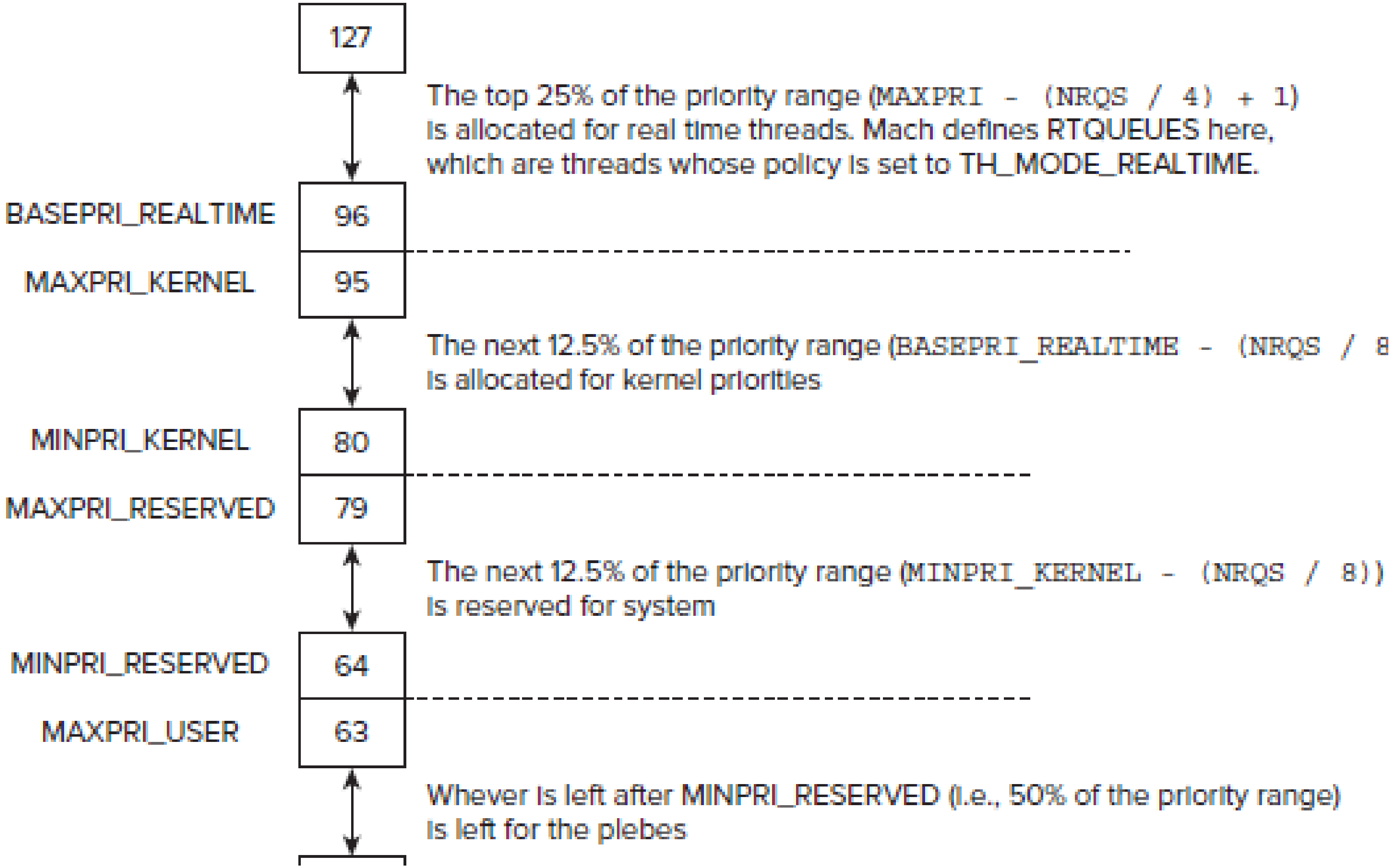
Tasks

- On its own, a task has no life. Its purpose is to serve as a container of one or more threads.
- The threads in a task are maintained in the threads member, which is a queue containing `thread_count` threads.

The Mach Priority Ranges



The Mach Priority Ranges



Indian Institute of Technology Kharagpur

Priority Shifts

- Mach dynamically tweaks the priorities of each thread during runtime, to accommodate for the thread's CPU usage, and overall system load.

- Threads can thus "drift" in their priority bands, decreasing in priority when using the CPU too much, and increasing in priority if not getting enough CPU.

Priority Shifts

- The traditional scheduler uses a function (`update_priority`) to update dynamically the priority of each thread. Gets called when the time quanta ends
  - Defined in `osfmk/kern/priority.c` (https://github.com/apple/darwin-xnu/blob/main/osfmk/kern/priority.c#L627)
- The macro toggles the thread priority by subtracting its calculated `sched_usage` (the amount of CPU used by this thread in the previous tick window), shifted by a `pri_shift` value.
- The `pri_shift` value is derived from the global `sched_pri_shift`, which is updated by the scheduler at every tick window as part of the system load calculation in `compute_averages` (`osfmk/kern/sched_average.c`).

Priority Shifts

- Subtracting the CPU usage delta effectively penalizes those threads with high CPU usage (positive usage delta detracts from priority) and rewards those of low CPU usage (negative usage delta adds to priority).
- To make sure the thread’s CPU usage doesn’t accrue to the point where the penalty is lethal, the `update_priority` function gradually ages CPU usage.
- It makes use of a `sched_decay_shifts` structure, to simulate the exponential decay of the CPU usage by a factor of (5/8)*n, defined in the `osfmk/kern/priority.c`
- By using the pre-computed shift values, the computation can be speed up, expressed in terms of bit shifts and additions, which take less time than multiplication.
- Recompute priority based on the CPU usage: https://github.com/apple/darwin-xnu/blob/main/osfmk/kern/priority.c#L544

The `sched_decay_shifts` structure in `osfmk/kern/priority.c`

```
/*
* Define shifts for simulating (5/8) ** n
*
* Shift structures for holding update shifts. Actual computation
* is usage = (usage >> shift1) +/- (usage >> abs(shift2)) where the
* +/- is determined by the sign of shift 2.
*/
struct shift_data {
        int shift1;
        int shift2;
};
// The shift data at index i provides the approximation of (5/8)i
#define SCHED_DECAY_TICKS 32
static struct shift_data sched_decay_shifts[SCHED_DECAY_TICKS] = {
{1,1},{1,3},{1,-3},{2,-7},{3,5},{3,-5},{4,-8},{5,7},
{5,-7},{6,-10},{7,10},{7,-9},{8,-11},{9,12},{9,-11},{10,-13},
{11,14},{11,-13},{12,-15},{13,17},{13,-15},{14,-17},{15,19},{16,18},
{16,-19},{17,22},{18,20},{18,-20},{19,26},{20,22},{20,-22},{21,-27}
};
```

Scheduling Algorithms

- Mach's thread scheduling is highly extensible, and actually allows changing the algorithms used for thread scheduling (`osfmk/kern/sched_prim.h`)

Supported schedulers in Mach	
KSCHED… CONSTANT (STRING)	**USED FOR**
SCHED_TRADITIONAL	Traditional (default)
SCHED_PROTO	Global runqueue based scheduler
SCHED_GRRR	Group Ratio Round Robin
SCHED_MULTIQ	Traditional multi-queue ready queue scheduler
SCHED_CLUTCH	Schedule group of threads
SCHED_EDGE	Better control over various QoS buckets

SCHED_CLUTCH - Schedule group of threads

- The clutch scheduler schedules groups of threads instead of individual threads.

- Breaking away from the traditional single-tier scheduling model, it implements a hierarchical scheduler which makes optimal decisions at various thread grouping levels.

- The hierarchical scheduler currently has 3 levels:
  - Scheduling Bucket Level
  - Thread Group Level
  - Thread Level

SCHED_CLUTCH - Schedule group of threads

- Clutch ordering based on thread group flags (specified by the thread grouping mechanism).
- These properties define a thread group specific priority boost.
- The current implementation gives a slight boost to HIGH & MED thread groups which effectively deprioritizes daemon thread groups which are marked "Efficient" on AMP systems.
- Bound threads are not managed in the clutch hierarchy.
- How to indicate if the thread should be in the hierarchy or not?

SCHED_CLUTCH - Schedule group of threads

- The clutch scheduler organizes the threads based on the thread group and the scheduling bucket. The Buckets are :

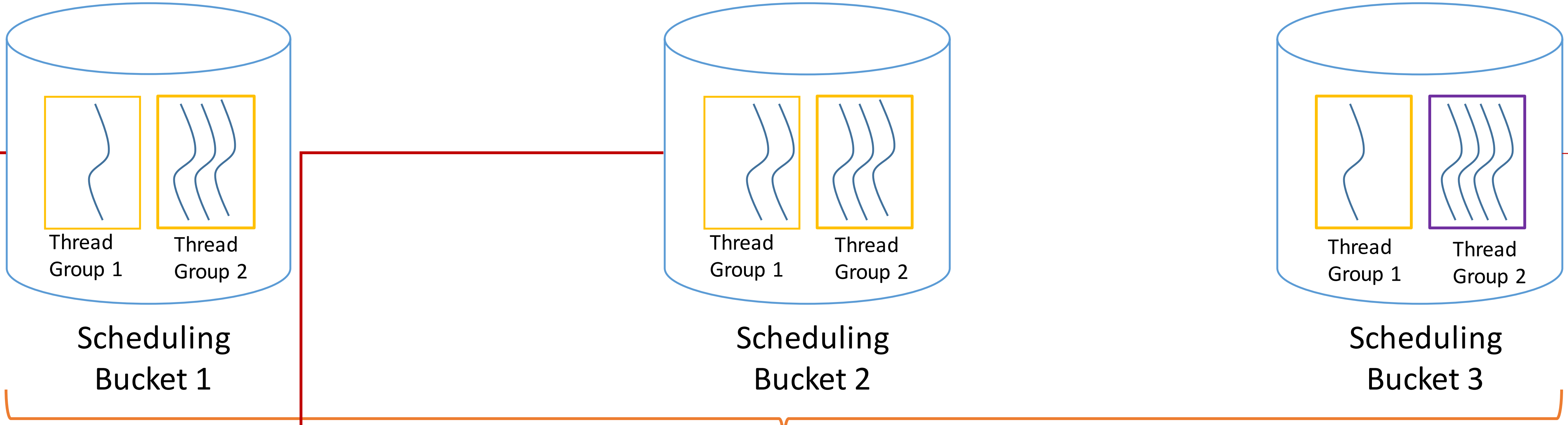
```
    TH_BUCKET_FIXPRI = 0,    /* Fixed-priority */
    TH_BUCKET_SHARE_FG,    /* Timeshare thread above BASEPRI_DEFAULT */
    /* Timeshare thread between BASEPRI_USER_INITIATED and BASEPRI_DEFAULT
*/
    TH_BUCKET_SHARE_IN
     /* Timeshare thread between BASEPRI_DEFAULT and BASEPRI_UTILITY */
    TH_BUCKET_SHARE_DF
    /* Timeshare thread between BASEPRI_UTILITY and MAXPRI_THROTTLE */
    TH_BUCKET_SHARE_UT
     /* Timeshare thread between MAXPRI_THROTTLE and MINPRI */
    TH_BUCKET_SHARE_BG
    TH_BUCKET_RUN,                          /* All runnable threads */
    TH_BUCKET_SCHED_MAX = TH_BUCKET_RUN,    /* Maximum schedulable buckets
*/
    TH_BUCKET_MAX,
```

SCHED_CLUTCH - Schedule group of threads

- In the clutch scheduler, the threads are maintained in runqs at the `clutch_bucket` level (`clutch_bucket` defines a unique thread group and scheduling bucket pair). The thread is linked via a couple of linkages in the clutch bucket:
  - A stable priority queue linkage which is the main runqueue (based on sched_pri) for the clutch bucket
  - A regular priority queue linkage which is based on thread's base/promoted pri (used for clutch bucket priority calculation)
  - A queue linkage used for timesharing operations of threads at the scheduler tick

- Since the clutch scheduler organizes threads based on the thread group and the scheduling bucket, it is important to not mix threads from multiple priority bands into the same bucket. To achieve that, in the clutch bucket world, there is a scheduling bucket per QoS effectively.

SCHED_CLUTCH - Scheduling Bucket Level



Decides which class of thread/root bucket should be picked for execution.

Earliest Deadline First (EDF) algorithm

These are scheduling buckets per thread maintained by kernel, and defined based on the base/scheduling priority of the threads.

All runnable threads with the same scheduling bucket are represented by a single entry in a priority queue which is ordered by the bucket's deadline.

SCHED_CLUTCH - Scheduling Bucket Level

- These entries are known as **root buckets** throughout the implementation.
- The goal of this level is to provide low latency access to the CPU for high QoS classes while ensuring starvation avoidance for the low QoS classes.
- The bucket selection algorithm simply selects the root bucket with the earliest deadline in the priority queue.
- The deadline for a root bucket is calculated based on its first-runnable timestamp and its **Worst Case Execution Latency (WCEL)** value which is pre-defined for each bucket.
- The WCEL values are picked based on the decay curve followed by the Mach timesharing algorithm to allow the system to function similar to the existing scheduler from a higher level perspective.

SCHED_CLUTCH - Scheduling Bucket Level

```
static uint32_t sched_clutch_root_bucket_wcel_us[TH_BUCKET_SCHED_MAX] = {
        SCHED_CLUTCH_INVALID_TIME_32,                   /* FIXPRI */
        0,                                              /* FG */
        37500,                                          /* IN (37.5ms) */
        75000,                                          /* DF (75ms) */
        150000,                                         /* UT (150ms) */
        250000                                          /* BG (250ms) */
};
```

SCHED_CLUTCH - Scheduling Bucket Level

- Whenever a root bucket transitions from non-runnable to runnable, its deadline is set to (now + WCEL[bucket]).

- This ensures that the bucket would be scheduled at WCEL[bucket] even in a heavily loaded system.

- Once the root bucket is picked for execution, its deadline is pushed by WCEL[bucket] into the future.

- For better performance, the bucket level scheduler implements a **root bucket warp mechanism.** Each bucket is provided a warp value which is refreshed whenever the bucket is selected due to its deadline expiring.

SCHED_CLUTCH - Scheduling Bucket Level

```
static uint32_t sched_clutch_root_bucket_warp_us[TH_BUCKET_SCHED_MAX] = {
        SCHED_CLUTCH_INVALID_TIME_32,                 /* FIXPRI */
        8000,                                         /* FG (8ms)*/
        4000,                                         /* IN (4ms) */
        2000,                                         /* DF (2ms) */
        1000,                                         /* UT (1ms) */
        0                                             /* BG (0ms) */
};
```

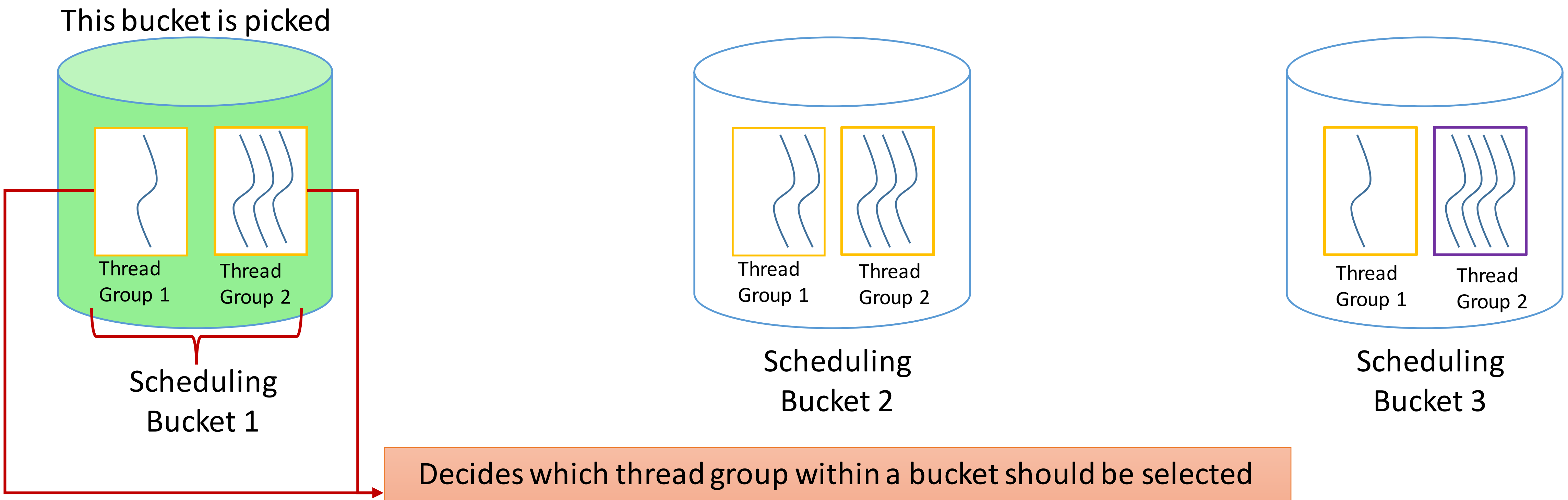
SCHED_CLUTCH - Scheduling Bucket Level

- The root bucket selection logic finds the earliest deadline bucket and then checks if there are any higher (in natural priority order) buckets that have warp remaining.
- If there is such a higher bucket, it would select that bucket and effectively open a warp window.
- During this warp window the scheduler would continue to select this warping bucket over lower priority buckets. Once the warping bucket is drained or the warp window expires, the scheduler goes back to scheduling buckets in deadline order.
- This mechanism provides a bounded advantage to higher level buckets to allow them to remain responsive in the presence of bursty workloads.
- The scheduling bucket level also maintains a bitmap of runnable root buckets to allow quick checks for empty hierarchy and root level priority calculation.

Earliest Deadline First (EDF) algorithm

- The EDF algorithm is the best choice for this level due to the following reasons:
  - Deadline based scheduling allows the scheduler to define strict bounds on worst case execution latencies for all scheduling buckets.
  - The EDF algorithm is dynamic based on bucket runnability and selection. Since all deadline updates are computationally cheap, the algorithm can maintain up-to-date information without measurable overhead.
  - It achieves the goals of maintaining low scheduling latency for high buckets and starvation avoidance for low buckets efficiently.
  - Since the bucket level scheduler deals with a fixed small number of runnable buckets in the worst case, it is easy to configure in terms of defining deadlines, warps etc.

SCHED_CLUTCH - Thread Group Level



Decides which thread group within a bucket should be selected next for execution.

Thread groups are a collection of threads working on behalf of a specific workload.

Each thread group with runnable threads within a bucket is represented as an entry at this level.

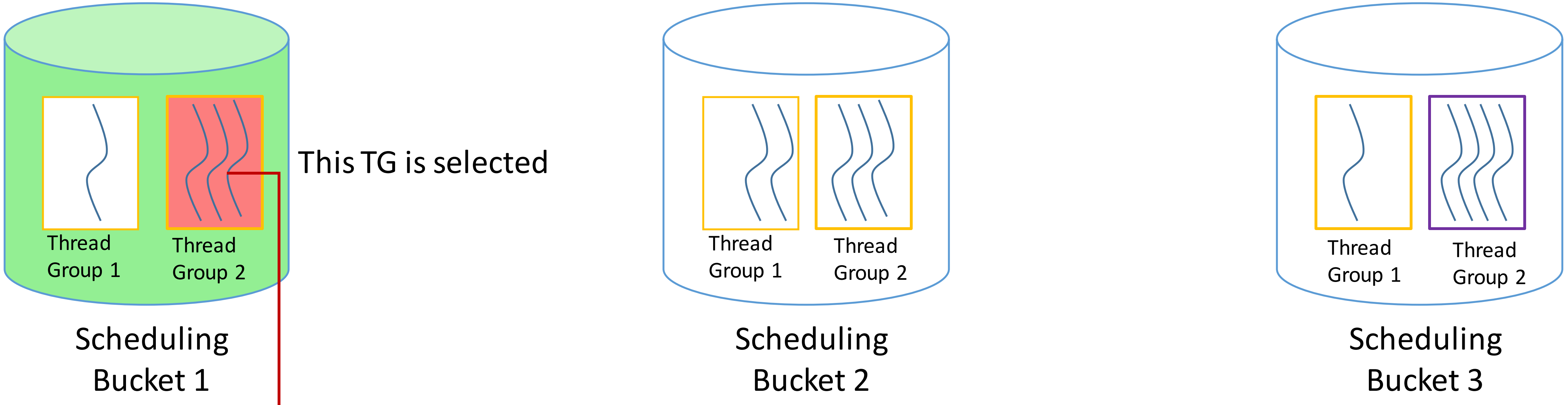
SCHED_CLUTCH - Thread Group Level

- These entries are known as **clutch buckets** throughout the implementation.
- The goal of this level is to share the CPU among various user workloads with preference to interactive applications over compute-intensive batch workloads.
- The clutch bucket selection algorithm simply selects the clutch bucket with the highest priority in the priority queue.
- The priority calculation for the clutch buckets is based on the following factors:
    - **Highest runnable thread in the clutch bucket**:
    - **Interactivity score**
    - **Thread Group Type**

SCHED_CLUTCH - Interactivity score based algorithm

- The interactivity score-based algorithm is well suited for this level due to the following reasons:
    - It allows for a fair sharing of CPU among thread groups based on their recent behavior. Since the algorithm only looks at recent CPU usage history, it also adapts to changing behavior quickly.
    - Since the priority calculation is fairly cheap, the scheduler can maintain up-to-date information about all thread groups which leads to more optimal decisions.
    - Thread groups provide a convenient abstraction for groups of threads working together for a user workload.
    - Scheduling decisions on this abstraction allows the system to make interesting choices such as preferring Apps over daemons which is typically better for system responsiveness.

SCHED_CLUTCH - Thread Level

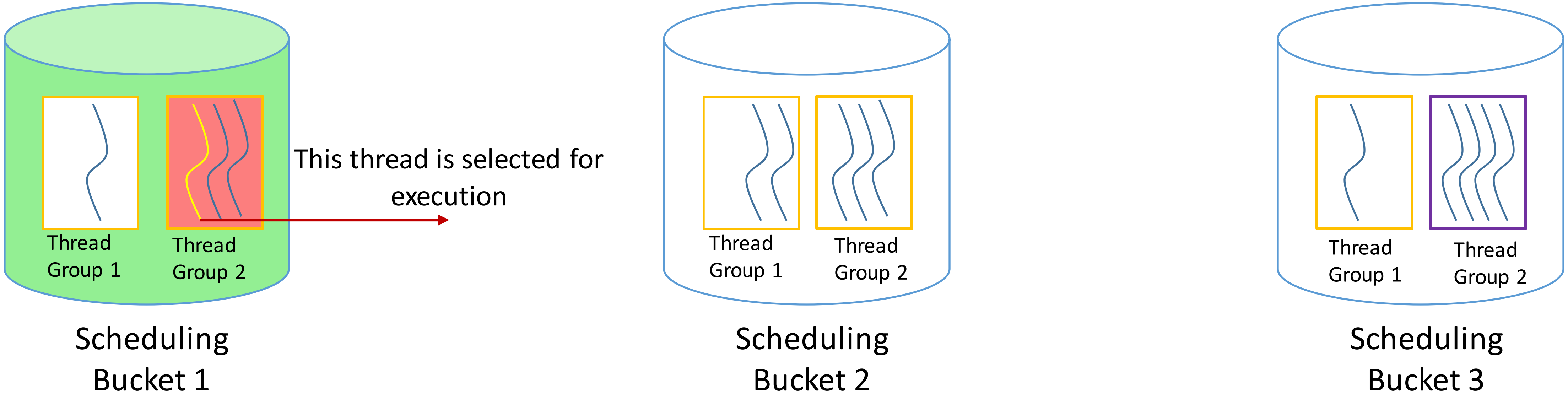


Decides which thread within a clutch bucket should be selected next for execution.

Each runnable thread in the clutch bucket is represented as an entry in a runqueue which is organized based on the schedpri of threads.

The thread selection algorithm simply selects the highest priority thread in the runqueue.

SCHED_CLUTCH - Thread Level



The schedpri calculation for the threads is based on the traditional Mach scheduling algorithm which uses load & CPU usage to decay priority for a thread.

The thread decay model is more suited at this level as compared to the global scheduler because the load calculation only accounts for threads in the same clutch bucket.

Since all threads in the same clutch bucket belong to the same thread group and scheduling bucket, this algorithm provides quick CPU access for latency sensitive threads within the clutch bucket without impacting other non-related threads in the system.

Scheduler Priority Calculations - Thread Level Implementation

- The thread level scheduler implements the Mach timesharing algorithm to decide which thread within the clutch bucket should be selected next for execution.

- All runnable threads in a clutch bucket are inserted into the runqueue based on the **`schedpri`**.

- The scheduler calculates the **`schedpri`** of the threads in a clutch bucket based on the number of runnable threads in the clutch bucket and the CPU usage of individual threads.

Scheduler Priority Calculations - Thread Level Implementation

- The load information is updated at every scheduler tick and the threads use this information for priority decay calculation as they use CPU.

- **The priority decay algorithm** attempts to reward bursty interactive threads and penalize CPU intensive threads. Once a thread is selected for running, it is assigned a quantum which is based on the scheduling bucket it belongs to.

- The quanta for various buckets are defined statically as:

```
static uint32_t sched_clutch_thread_quantum_us[TH_BUCKET_SCHED_MAX] = {
    10000, /* FIXPRI (10ms) */
    10000, /* FG (10ms) */
    8000, /* IN (8ms) */
    6000, /* DF (6ms) */
    4000, /* UT (4ms) */
    2000 /* BG (2ms) */ };
```

Scheduler Priority Calculations - Root Priority Calculation

- The scheduler maintains a root level priority for the hierarchy to make decisions regarding preemptions and thread selection.
- The root priority is updated as threads are inserted/removed from the hierarchy.
- The root level also maintains the urgency bits to help with preemption decisions.
- Since the root level priority/urgency is used for preemption decisions, it is based on the threads in the hierarchy and is calculated as shown in the next slide.

Scheduler Priority Calculations - Root Priority Calculation

- Root Priority Calculation:

  1. If AboveUI bucket (buckets with threads for user interface interactions) is runnable,
  2. Compare priority of AboveUI highest clutch bucket (CBUI) with Timeshare FG highest clutch bucket (CBFG)
  3. If pri(CBUI) >= pri(CBFG), select CBUI
  4. Otherwise find the (non-AboveUI) highest priority root bucket that is runnable and select its highest clutch bucket
  5. Find the highest priority (promoted or base pri) thread within that clutch bucket and assign that as root priority

Scheduler Priority Calculations - Clutch Bucket Priority Calculation

- As mentioned earlier, the priority value of a clutch bucket is calculated based on the highest runnable thread, interactivity score and the thread group type.

1. Find the highest runnable thread (promoted or basepri) in the clutch bucket (maxpri)
2. Check if the thread group for this clutch bucket is marked *Efficient*.
3. If not, assign a positive boost value (clutch_boost)
4. Calculate the ratio of CPU blocked and CPU used for the clutch bucket.
5. If blocked > used, assign a score (interactivity_score) in the higher range.
6. Else, assign a score (interactivity_score) in the lower range.
7. clutch-bucket priority = maxpri + clutch_boost + interactivity_score

Scheduler Priority Calculations - Thread Priority Calculation

- The thread priority calculation is based on the Mach timesharing algorithm.

1. At every scheduler tick, snapshot the load for the clutch bucket
2. Use the load value to calculate the priority shift values for all threads in the clutch bucket
3. thread priority = base priority - (thread CPU usage >> priority shift)

SCHED_GRRR – Group Ratio Round Robin Scheduler

Group Ratio Round-Robin: O(1) Proportional Share Scheduling for Uniprocessor and Multiprocessor Systems

Bogdan Caprita, Wong Chun Chan, Jason Nieh, Clifford Stein, and Haoqiang Zheng*
Department of Computer Science
Columbia University
Email: {bc2008, wc164, nieh, cliff, hzheng}@cs.columbia.edu

Abstract

We present Group Ratio Round-Robin (GR^3), the first proportional share scheduler that combines accurate proportional fairness scheduling behavior with $O(1)$ scheduling overhead on both uniprocessor and multiprocessor systems. GR^3 uses a simple grouping strategy to organize clients into groups of similar processor allocations which can be more easily scheduled. Using this strategy, GR^3 combines the benefits of low overhead round-robin execution with a novel ratio-based scheduling algorithm. GR^3 introduces a novel frontlog mechanism and weight readjustment algorithm to operate effectively on multiprocessors. GR^3 provides fairness within a constant factor of the ideal generalized processor sharing model for client weights with a fixed upper bound and preserves its fairness properties on multiprocessor systems. We have implemented GR^3 in Linux and measured its performance. Our experimental results show that GR^3 provides much lower scheduling overhead and much better scheduling accuracy than other schedulers commonly used in research and practice.

work has been done to provide proportional share scheduling on multiprocessor systems, which are increasingly common especially in small-scale configurations with two or four processors. Over the years, a number of scheduling mechanisms have been proposed, and much progress has been made. However, previous mechanisms have either superconstant overhead or less-than-ideal fairness properties.

We introduce Group Ratio Round-Robin (GR^3), the first proportional share scheduler that provides constant fairness bounds on proportional sharing accuracy with $O(1)$ scheduling overhead for both uniprocessor and small-scale multiprocessor systems. In designing GR^3, we observed that accurate, low-overhead proportional sharing is easy to achieve when scheduling a set of clients with equal processor allocations, but is harder to do when clients require very different allocations. Based on this observation, GR^3 uses a simple client grouping strategy to organize clients into groups of similar processor allocations which can be more easily scheduled. Using this grouping strategy, GR^3 combines the benefits of low overhead round-robin execu-

USENIX ATC 2005

SCHED_GRRR - Group Ratio Round Robin

- The Group Ratio Round-Robin (GR^3) scheduler is the first proportional share scheduler [1] that combines accurate proportional fairness scheduling[2] behavior with O(1) scheduling overhead on both uniprocessor and multiprocessor systems, i.e. it provides constant fairness bounds on proportional sharing accuracy with O(1) scheduling overhead.

- GR^3 uses a simple **grouping strategy** to organize clients into groups of similar processor allocations which can be more easily scheduled.
- Why to use grouping strategy?
  - Observations: Accurate, low-overhead proportional sharing is easy to achieve when scheduling a set of clients with equal processor allocations, but is harder to do when clients require very different allocations

[1] In a proportional share algorithm every job has a weight, and jobs receive a share of the available resources proportional to the weight of every job.
[2] Proportional-fair scheduling is a compromise-based scheduling algorithm. It is based upon maintaining a balance between two competing interests.

Interrupt-driven Scheduling

- Like other contemporary OS, Mach OS also uses the hardware interrupts to usurp the control of the CPU from the existing thread.
- Interrupts forces the CPU to “drop everything” on interrupt and long jump to the interrupt handler (also known as the interrupt service routine, or ISR).
- Apart from asynchronous interrupt, XNU also provides an interrupt that can be triggered in a given time frame for a predictable interrupt source, namely, real time clock, or rtclock.
- This clock is hardware dependent — the Intel architecture uses the local CPU’s programmable interrupt controller (APIC) for this purpose — and can be configured by the kernel to generate an interrupt after a given number of cycles.
    - This interrupt source is typically referred to as the ***Timer Interrupt***

Interrupt-driven Scheduling

- Older versions of XNU triggered the Timer Interrupt at a fixed number of times per second, a value referred to as hz.

- This value is globally defined in the BSD portion of the kernel, in bsd/kern/clock.c (now deprecated)

Indian Institute of Technology Kharagpur

Issue with Fixed interval in Timer Interrupt

- A timer interrupting the kernel at a fixed interval will cause predictable, but extraneous interrupts.

- Too high a value of hz implies too many unnecessary interrupts.

- On the other hand, too low a value would mean the system is less responsive, as sub-hz delays would only be achievable by a tight loop.

- The old `hertz_tick()` function used in previous versions of OS X is still present, but unused and conditionally compiled only if XNU is compiled with profiling.

Indian Institute of Technology Kharagpur

Solution - A *tick-less* kernel

- A tick-less kernel, in which on every timer interrupt, the timer is *reset* to schedule the next interrupt only when the scheduler seems it necessary.

- This means that, on every timer interrupt, the interrupt handler must make a (very quick) pass over the list of pending deadlines, which are primarily sleep timeouts set by threads, act on them, if necessary, and schedule the next timer interrupt accordingly.

- More processing in each timer interrupt is well worth the savings in spurious interrupts, and the processing can be kept to a minimum by keeping track of only the most exigent deadline.

Indian Institute of Technology Kharagpur