

macOS 内核结构简介

李亮

holly.lee@gmail.com

Nov, 2016

当我们谈论操作系统时，我们在谈论什么

- 核心
- 基本用户程序
- GUI
- 开发者: syscall, loader, executable format, ...

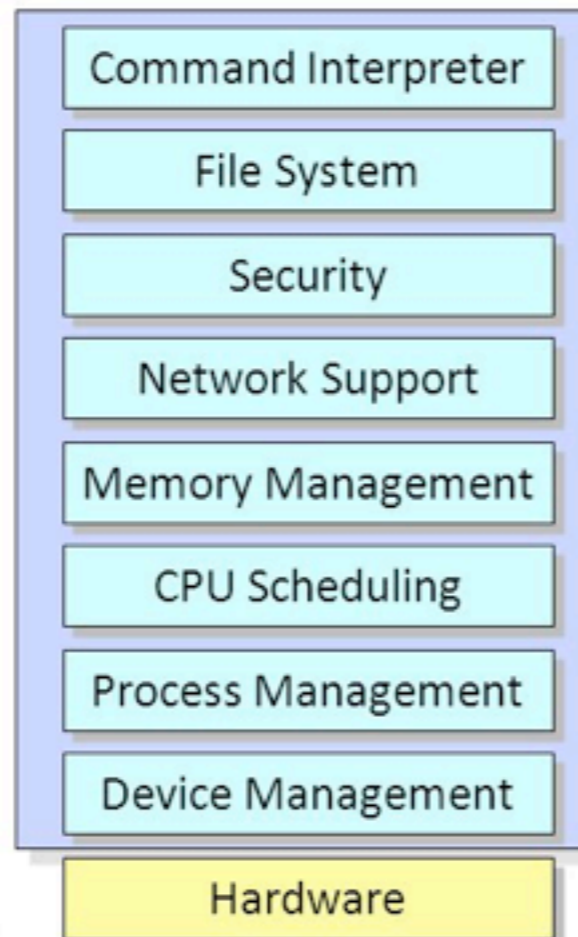
内核结构的类型

- 微内核: Mach
- 一体化内核: Linux
- 混合式: Mac OS X, Windows

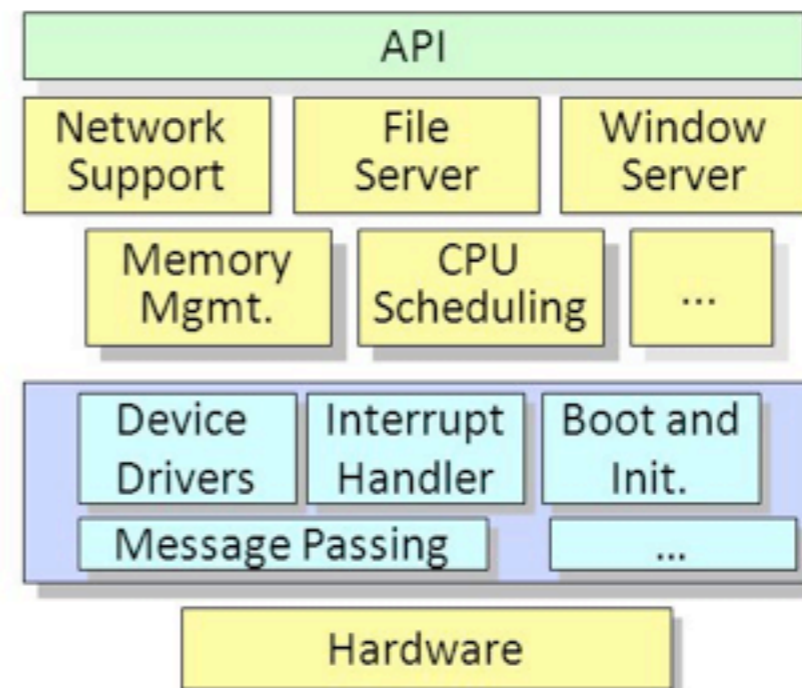
微内核 vs 一体化内核

Operating System Structures

- Monolithic OS (e.g., Unix)

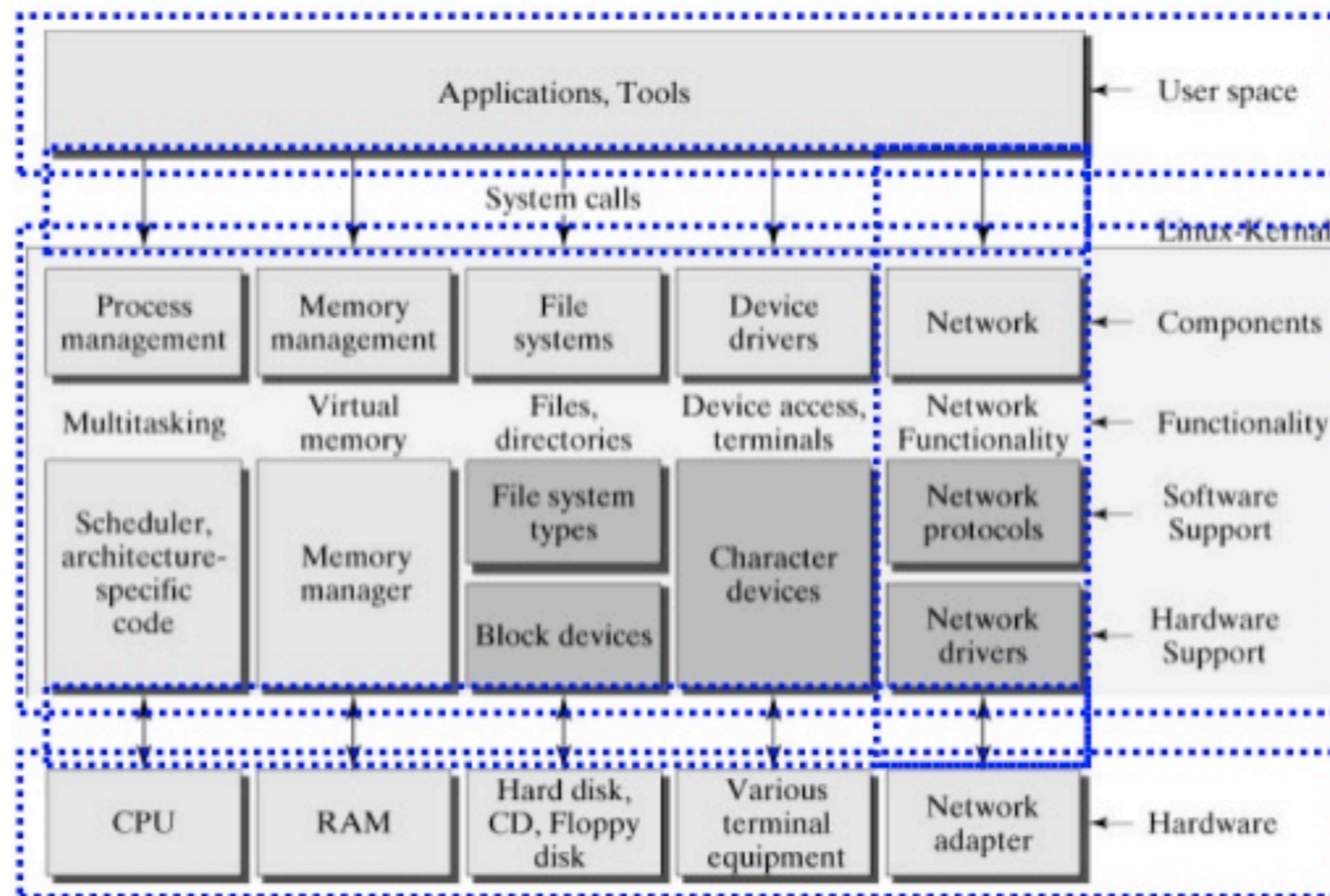


- Micro-kernel OS (e.g., Mach, Exokernel, ...)

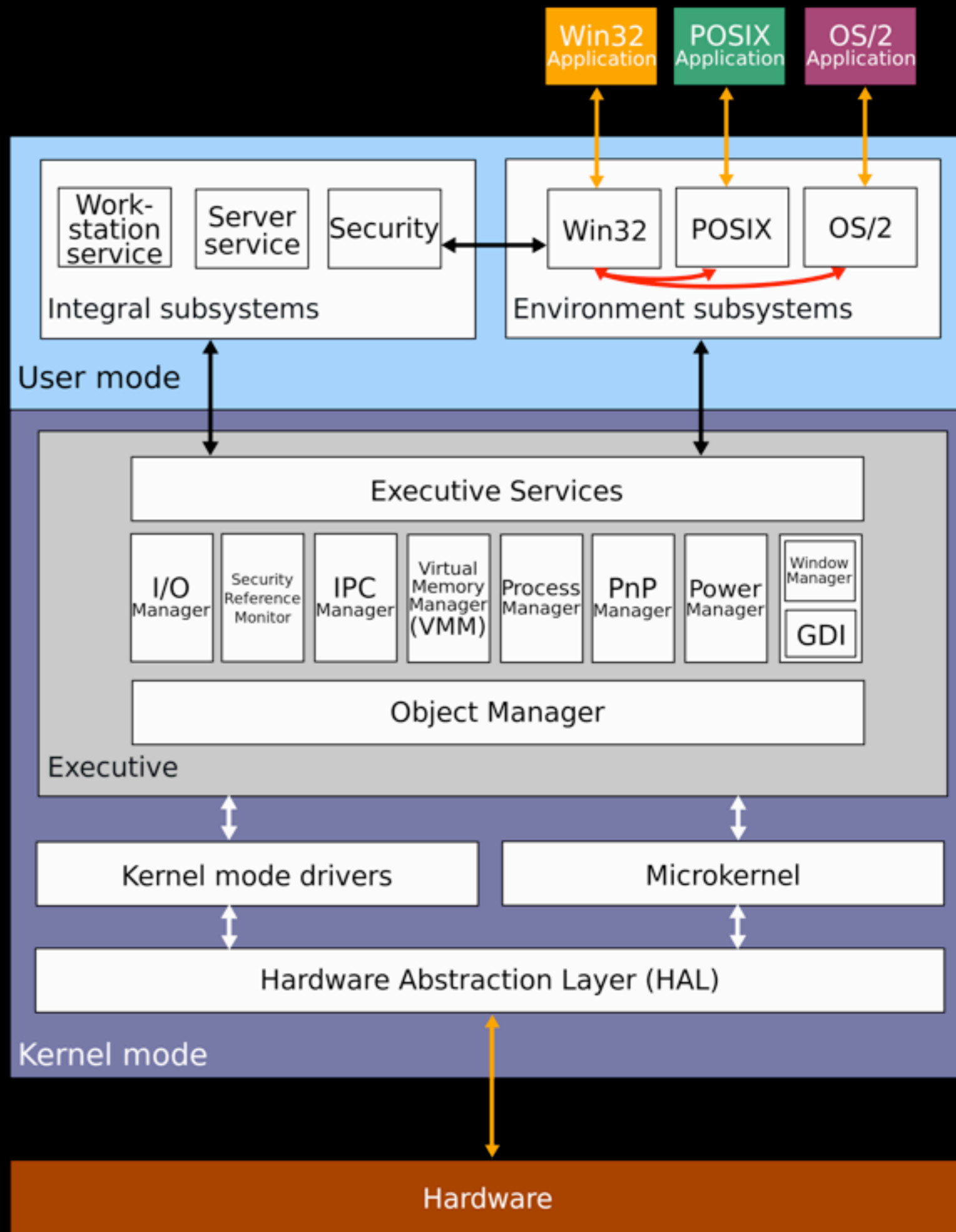


一体化内核: Linux

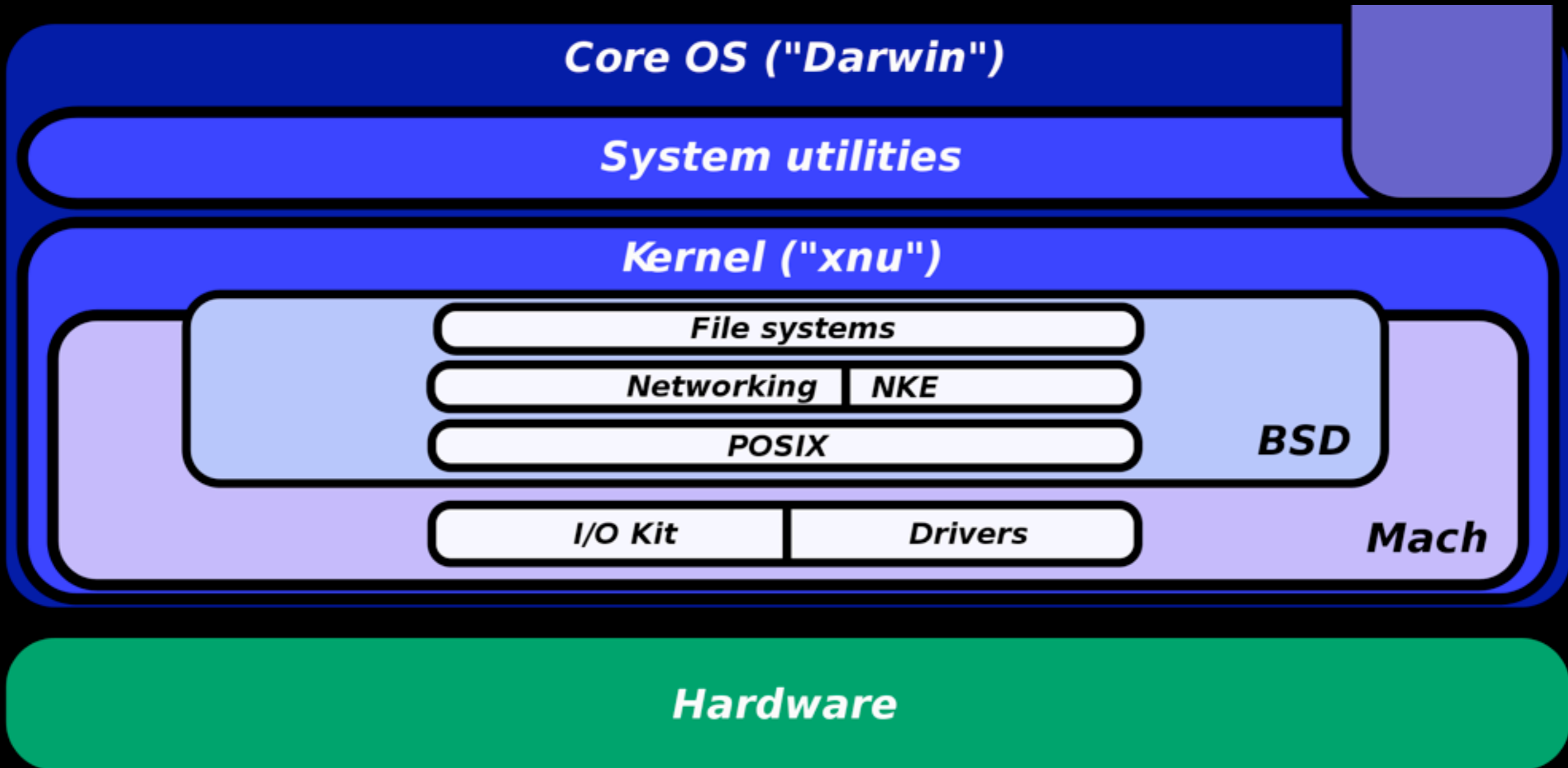
Linux Kernel Structure



混合内核: Windows



OS X 内核结构图



Mach in General

- Mach. 面向对象的设计思路. 尽可能小的核心, 外围一堆 server 提供需要的外围功能. 利用消息进行功能的请求与完成.
- 纯粹的 Mach. 在以用户空间进程的形式来实现各个 Server. 消息即 IPC.
- IPC 导致的性能问题. Context switch, memory mapping, cache missing, etc.

Mach in OS X


- 演变. 从严格的到不那么严格的.
- Mach 部分成为了提供最核心的, 必须的操作系统服务的部分.
- thread Object, task Object, scheduler, clock object, VMM, processor object, synchronization object.
- **IPC**: mach messages/ports. 依然是很重要的部分. 大量用于用户进程跟 kernel 之间的通信, 用户进程之间的通信, 以及内核代码中需要访问另一进程的资源时.

Mach in OS X

- 与机器本身相关的: host object, processor object, processor set object, clock object.
- 调度相关的: task, thread, exception handling.
- 同步相关: mutex, spinlock, semaphore, rwlock.
- 内存管理.

Mach: task & thread

- thread: 最小调度单位.
- task: 容器. 包含 threads, memory space, etc.
- task 概念与进程概念. BSD 层的进程与 task 的一一对应关系.
- thread. BSD 层的线程与 thread 的对应.
- 有用户空间可访问的相关 API. 返回当前, 信息, 创建, 销毁, 取得 port, etc.

| Process Name | % CPU ▾ | CPU Time | Threads | Idle Wake Ups | PID | User |
|---|---------|------------|---------|---------------|-------|------|
| ▼ kernel_task | 3.0 | 2:44:24.39 | 139 | 1,417 | 0 | root |
| ▼ launchd | 0.0 | 29:55.66 | 4 | 2 | 1 | root |
| ▼  Firefox | 11.6 | 5:17:26.74 | 63 | 401 | 10720 | lee |

Mach: 调度

- 通用的. 抢占式, 优先级, 运行队列, 等待队列, CPU Affinity.
- 显式抢占和隐式抢占. 显式: 因为等待资源而阻塞; 隐式: 时间片用完.
- 特有: Handoff, 主动放弃并指定交给哪个线程继续, 特别符合 Mach 的消息传递机制
- 特有: Continuation. 线程指定在下次唤醒后执行什么而非继续. 省时间和栈空间. 只在显式抢占.
- 特有: Asynchronous Software Trap. 类似 Linux 的 SoftIRQ, 在 trap 从内核返回到用户空间时执行.

Mach: exception handling

- 利用 Mach message.
- 处理器 exception handler: mach_msg() 发送一个 mach message 到 exception port. 并等待回应.
- exception port: 三级. thread, task and host
- 实际的真正的 exception 处理在 BSD 层, 在那里被转化成 Unix signal. BSD 注册了一个 host exception port.
- Crash reporter (launchd 注册 task exception port 并被所有进程继承) / Debugger (为 debugger 注册 exception port) 都是利用这个机制注册相应的 handler 来完成其任务的.

Mach: 内存管理

- 物理内存的管理. pmap. 页表, 页表项, TLB
- 虚拟内存的管理. vm_map, vm_map_entry, vm_object, vm_page.
- 虚拟内存的交换. pager. 允许多个 pager 的存在. 多种不同的 back store 以及不同的读写方式. pageout kernel threads 管理 page list 和交换动作.
- Freezer. (iOS)
- 用户空间 API. 进程虚拟内存空间的访问.

BSD 层

在 Mach 层提供的基础服务之上, 将相关的部分从 FreeBSD 移植了过来.

VFS, Networking, Process management, User/Group, ...

POSIX 标准

BSD 层: 进程和线程

- BSD 的 Process 跟 Mach 的 task 是一一对应的

```
struct task { void * bsd_info; } <-> struct proc { void * task; }
```

- 增加了诸如 fd list, uid/gid, 进程关系, tty, signal 处理等相关的数据,

- BSD 的 Thread 跟 Mach 的 thread 是一一对应的

```
struct thread { void * uthread; }  
                <->  
struct uthread { struct vfs_thread { thread_t vc_thread; } * uu_context; }
```

- Mach 调用: task_for_pid().
- 调试. ptrace() 系统调用.

BSD 层: 可执行文件

- 可执行文件格式以及 Loader 在内核的部分都是位于 BSD 层的.
- Mach-O 格式. 支持 Fat binary. 由多个 Load Command 组成. 比如 LC_SEGMENT 包含了实际要映射到内存里的代码/数据段, LC_CODESIGNATURE 包含了签名信息等等. 对应于 ELF 格式中的各个 section, PE 里的各个 table.
- dyld. OS X 的 loader.

BSD 层: 直接移植的

- VFS. 各种文件系统. disk image. APFS
- 整个网络栈.
- User/Group 机制.

BSD 层: POSIX

通过这一层, 实现了大部分 POSIX 标准, 使得程序的移植利用, 变得很方便, 同时又不失去随时替换下层 Mach 部分的可能性.

IOKit

Driver framework

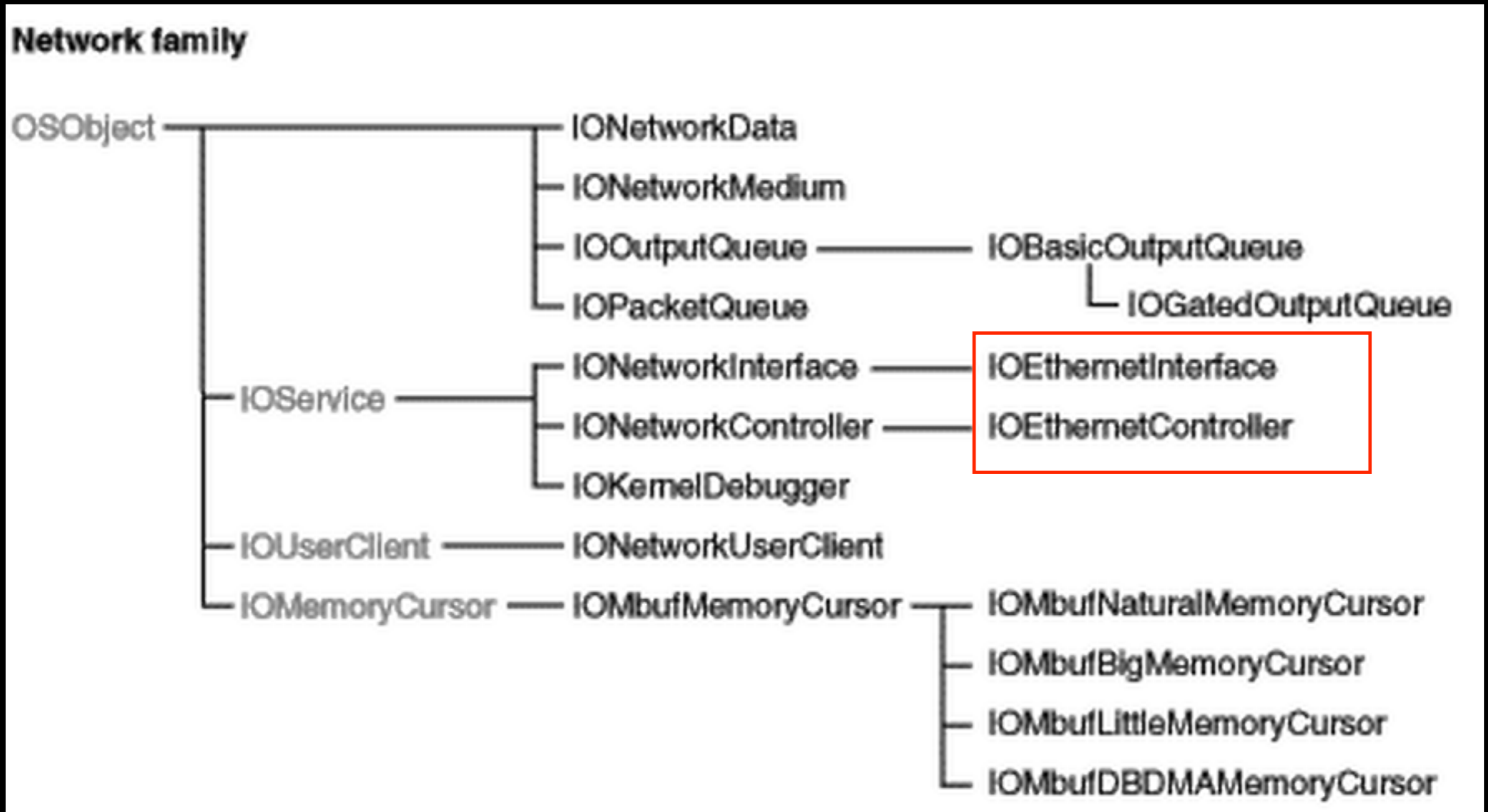
C++

libkern: 基础类库

IOKit: IORegistry

- 系统中有一个 IORegistry 登记了所有的外部设备
- IOKit driver 将符合自身的设备加入 IORegistry
- IOKit driver 也利用 IORegistry 的信息来匹配发现的外设.
- 用户空间有 API. 也有命令行可以查看: ioreg. GUI: IORegistryExplorer.

IOKit: 网络部分示例



OS X 与 iOS

- 绝大部分是一样的.
- iOS: Jetsam
- iOS: Sandbox
- iOS: Entitlement
- iOS: Encrypted image

参考资料

- <http://opensource.apple.com>
- *Mac OS X Internals* by Amit Singh
- *Mac OS X and iOS Internals: To Apple Core*, by Jonathan Levin.

谢谢大家

李亮

holly.lee@gmail.com

Weibo: @holly_lee

Twitter: @opensky2

blog.freecoder.org