Monolithic Kernels vs Microkernels

- Idea of microkernel:
  - Flexible, minimal platform
  - Mechanisms, not policies
  - Goes back to Nucleus [Brinch Hansen, CACM’70]

Microkernel Evolution

First generation
- Eg Mach ['87]
  - ~180 syscalls
  - ~100 KLOC
  - ~100 µs IPC

Second generation
- Eg L4 ['95]
  - ~7 syscalls
  - ~10 KLOC
  - ~1 µs IPC

Third generation
- seL4 ['09]
  - ~3 syscalls
  - 9 KLOC
  - 0.2–1 µs IPC
2nd-Generation Microkernels

- 1st-generation kernels (Mach, Chorus) were a failure
  - Complex, inflexible, slow
- L4 was first 2G microkernel [Liedtke, SOSP'93, SOSP'95]
  - Radical simplification & manual micro-optimisation
  - "A concept is tolerated inside the microkernel only if moving it outside the kernel, i.e. permitting competing implementations, would prevent the implementation of the system's required functionality."
  - High IPC performance
- Family of L4 kernels:
  - Original Liedtke (GMD) assembler kernel ('95)
  - Family of kernels developed by Dresden, UNSW/NICTA, Karlsruhe
  - Commercial clones (PikeOS, P4, CodeZero, ...)
  - Influenced commercial QNX ('82), Green Hills Integrity ('90s)
  - Generated NICTA startup Open Kernel Labs (OK Labs)
    - large-scale commercial deployment (multiple billions shipped)

Issues of 2G Microkernels

- L4 solved performance issue [Härtig et al, SOSP'97]
- Left a number of security issues unsolved
- Problem: ad-hoc approach to protection and resource management
  - Global thread name space \(\Rightarrow\) covert channels [Shapiro'03]
  - Threads as IPC targets \(\Rightarrow\) insufficient encapsulation
  - Single kernel memory pool \(\Rightarrow\) DoS attacks
  - Insufficient delegation of authority \(\Rightarrow\) limited flexibility, performance
- Addressed by seL4
  - Designed to support safety- and security-critical systems

seL4 Principles

- Single protection mechanism: capabilities
  - Except for time 😊
- All resource-management policy at user level
  - Painful to use
  - Need to provide standard memory-management library
    - Results in L4-like programming model
- Suitable for formal verification (proof of implementation correctness)
  - Attempted since '70s
  - Finally achieved by L4.verified project at NICTA [Klein et al, SOSP'09]

seL4 Concepts

- Capabilities (Caps)
  - mediate access
- Kernel objects:
  - Threads (thread-control blocks, TCBs)
  - Address spaces (page table objects, PDs, PTs)
  - IPC endpoints (EPs, AsyncEPs)
  - Capability spaces (CNodes)
  - Frames
  - Interrupt objects
  - Untyped memory
- System calls
  - Send, Wait (and variants)
  - Yield
Capabilities (Caps)

- Token representing privileges [Dennis & Van Horn, ’66]
  - Cap = “prima facie evidence of right to perform operation(s)”

- Object-specific ⇒ fine-grained access control
  - Cap identifies object ⇒ is an (opaque) object name
  - Leads to object-oriented API:
    \[
    \text{err} = \text{method}(\cap, \text{args});
    \]
    - Privilege check at invocation time

- Caps were used in microkernels before
  - KeyKOS (’85), Mach (’87)
  - EROS (’99): first well-performing cap system
  - OKL4 V2.1 (’08): first cap-based L4 kernel

seL4 Capabilities

- Stored in cap space (CSpace)
  - Kernel object made up of CNodes
    - each an array of cap “slots”

- Inaccessible to userland
  - But referred to by pointers into CSpace (slot addresses)
  - These CSpace addresses are called CPTRs

- Caps convey specific privilege (access rights)
  - Read, Write, Grant (cap transfer) [Yes, there should be Execute!]

Main operations on caps:

- \text{Invoke}: perform operation on object referred to by cap
  - Possible operations depend on object type
- \text{Copy/Mint/Grant}: create copy of cap with same/lesser privilege
- \text{Move/Mutate}: transfer to different address with same/lesser privilege
- \text{Delete}: invalidate slot
  - Only affects object if last cap is deleted
- \text{Revoke}: delete any derived (eg. copied or minted) caps

Inter-Process Communication (IPC)

- Fundamental microkernel operation
  - Kernel provides no services, only mechanisms
  - OS services provided by (protected) user-level server processes
  - Invoked by IPC

- seL4 IPC uses a handshake through endpoints:
  - Transfer points without storage capacity
  - Message must be transferred instantly
    - One partner may have to block
    - Single copy user → user by kernel

- Two endpoint types:
  - Synchronous (Endpoint) and asynchronous (AsyncEP)

Synchronous Endpoint

- Threads must rendez-vous for message transfer
  - One side blocks until the other is ready
  - Implicit synchronisation

- Message copied from sender’s to receiver’s message registers
  - Message is combination of caps and data words
    - presently max 121 words (484B, incl message “tag”)
**Asynchronous Endpoint**

- Avoids blocking
  - send OR’s cap badge to AEP’s data word
  - no caps can be sent
- Receiver can poll or wait
  - waiting returns and clears data word
  - polling just returns data word
- Similar to interrupt (with small payload, like interrupt mask)

**Sync Endpoints are Message Queues**

- EP has no sense of direction
- May queue senders or receivers
  - never both at the same time!
- Communication needs 2 EPs!

**Receiving from Sync and Async Endpoints**

Server with synchronous and asynchronous interface

- Example: file system
  - synchronous (RPC-style) client protocol
  - asynchronous notifications from driver
- Could have separate threads waiting on endpoints
  - forces multi-threaded server, concurrency control
- Alternative: allow single thread to wait on both EP types
  - Mechanism:
    - AsyncEP is bound to thread with BindAEP() syscall
    - thread waits on synchronous endpoint
    - async message delivered as if been waiting on AsyncEP

**Client-Server Communication**

- Asymmetric relationship:
  - Server widely accessible, clients not
  - How can server reply back to client (distinguish between them)?
- Client can pass (session) reply cap in first request
  - server needs to maintain session state
  - forces stateful server design
- seL4 solution: Kernel provides single-use reply cap
  - only for Call operation (Send+Wait)
  - allows server to reply to client
  - cannot be copied/minted/re-used but can be moved
  - one-shot (automatically destroyed after first use)
Call RPC Semantics

Client

Kernel

Server

Call(ep, …)

mint rep
deliver to server

process

Send(rep, …)

deliver to client

destroy rep

Identifying Clients

Stateful server serving multiple clients

• Must respond to correct client
  - Ensured by reply cap
• Must associate request with correct state
• Could use separate EP per client
  - Endpoints are lightweight (16 B)
  - But requires mechanism to wait on a set of EPs (like select)
• Instead, seL4 allows to individually mark ("badge") caps to same EP
  - Server provides individually badged caps to clients
  - Server tags client state with badge
  - Kernel delivers badge to receiver on invocation of badged caps

IPC Mechanics: Virtual Registers

• Like physical registers, virtual registers are thread state
  - Context-switched by kernel
  - Implemented as physical registers or thread-local memory location
• Message registers
  - Contain message transferred in IPC
  - Architecture-dependent subset mapped to physical registers
    • 5 on ARM, 3 on x86
    • Library interface hides details
  - 1st message register is special, contains message tag
• Reply cap
  - Overwritten by next receive!
  - Can move to CSpace with cspace_save_reply_cap()

IPC Message Format

Tag

Message

Caps (on Send)

Badges (on Receive)

CSpace reference for receiving caps (Receive only)

Note: Don’t need to deal with this explicitly for project
**Client-Server IPC Example**

**Server**

```c
seL4_Word addr = ut_alloc(seL4_EndpointBits);
err = cspace_ut_retype_addr(tcb_addr, seL4_EndpointObject, seL4_EndpointBits, cur_cspace, &ep_cap);
seL4_CPtr cap = cspace_mint_cap(dest, cur_cspace, ep_cap, seL4_all_rights, seL4_CapData_MakeBadge(0));
...
seL4_Word badge;
seL4_MessageInfo_t msg = seL4_Wait(ep, &badge);
seL4_CPtr slot = cspace_save_reply_cap(cur_cspace);
...
seL4_MessageInfo_t reply = seL4_MessageInfo_new(0, 0, 0, 0);
seL4_Send(slot, reply);
cspace_free_cslot(slot);
```

**Client**

```c
seL4_MessageInfo_t tag = seL4_MessageInfo_new(0, 0, 0, 1);
seL4_SetTag(tag);
seL4_SetMR(0, 1);
seL4_Call(server_c, tag);
```

**IPC Operations Summary**

- **Send** (ep_cap, ...), **Wait** (ep_cap, ...), **Wait** (aep_cap, ...)
  - blocking message passing
  - needs Write, Read permission, respectively
- **NBSend** (ep_cap, ...)
  - discard message if receiver isn’t ready
- **Call** (ep_cap, ...)
  - equivalent to **Send** (ep_cap, ...)+ reply-cap + **Wait** (ep_cap, ...)
- **Reply** (...)
  - equivalent to **Send** (rep_cap, ...)
- **ReplyWait** (ep_cap, ...)
  - equivalent to **Reply** (...)+ Wait (ep_cap, ...)
  - purely for efficiency of server operation
- **Notify** (aep_cap, ...), **Poll** (aep_cap, ...)
  - non-blocking send / check for message on AsyncEP

**Server Saving Reply Cap**

```c
seL4_Here cap = cspace_mint_cap(cspace, seL4_CapData_MakeBadge(0));
seL4_CPtr slot = cspace_save_reply_cap(cur_cspace);
seL4_MessageInfo_t reply = seL4_MessageInfo_new(0, 0, 0, 0);
seL4_Send(slot, reply);
cspace_free_cslot(slot);
```

**Derived Capabilities**

- **Badging** is an example of **capability derivation**
- The **Mint** operation creates a new, less powerful cap
  - Can add a badge
  - Can strip access rights
  - eg WR→R/O
- **Granting** transfers caps over an Endpoint
  - Delivers copy of sender’s cap(s) to receiver
    - reply caps are a special case of this
  - Sender needs Endpoint cap with Grant permission
  - Receiver needs Endpoint cap with Write permission
    - else Write permission is stripped from new cap
- **Retyping**
  - Fundamental operation of seL4 memory management
  - Details later…

---

©2011 Gernot Heiser UNSW/NICTA. Distributed under Creative Commons Attribution License
**seL4 System Calls**

- Notionally, seL4 has 6 syscalls:
  - `Yield()`: invokes scheduler
    - only syscall which doesn’t require a cap!
  - `Send()`, `Receive()` and 3 variants/combinations thereof
    - `Notify()` is actually not a separate syscall but same as `Send()`
    - This is why I earlier said "approximately 3 syscalls"

- All other kernel operations are invoked by "messaging"
  - Invoking `Send()`/`Receive()` on an object cap
  - Each object has a set of kernel protocols
    - operations encoded in message tag
    - parameters passed in message words
  - Mostly hidden behind "syscall" wrappers

**seL4 Memory Management Principles**

- Memory (and caps referring to it) is typed:
  - Untyped memory:
    - unused, free to Retype into something else
  - Frames:
    - (can be) mapped to address spaces, no kernel semantics
  - Rest: TCBs, address spaces, CNodes, EPs
    - used for specific kernel data structures

- After startup, kernel never allocates memory!
  - All remaining memory made Untyped, handed to initial address space
  - Space for kernel objects must be explicitly provided to kernel
    - Ensures strong resource isolation
  - Extremely powerful tool for shooting oneself in the foot!
    - We hide much of this behind the cspace and ut allocation libraries

**Capability Derivation**

- Copy, Mint, Mutate, Revoke are invoked on CNodes
  - CNode cap must provide appropriate rights
  - Copy takes a cap for destination
    - Allows copying of caps between CSpaces
    - Alternative to granting via IPC (if you have privilege to access Cspace!)

**Cspace Operations**

```c
extern seL4_CPtr cspace_copy_cap(cspace_t* dest, cspace_t* src, seL4_CPtr src_cap, seL4_CapRights rights);
extern seL4_CPtr cspace_mint_cap(cspace_t* dest, cspace_t* src, seL4_CPtr src_cap, seL4_CapRights rights, seL4_CapData badge);
extern seL4_CPtr cspace_move_cap(cspace_t* dest, cspace_t* src, seL4_CPtr src_cap);
extern cspace_err_t cspace_delete_cap(cspace_t* c, seL4_CPtr cap);
extern cspace_err_t cspace_revoke_cap(cspace_t* c, seL4_CPtr cap);
```
**cspace and ut libraries**

- **OS Personality**
- **System Calls**
- **Library Calls**
- **ut_alloc()**
- **ut_free()**
- **cspace_create()**
- **cspace_destroy()**

- Manages slab of Untyped
- Wraps messy Cspace tree & slot management
- Extend for own needs!

**seL4 Memory Management Approach**

- Resources fully delegated, allows autonomous operation
- Strong isolation, No shared kernel resources

**Global Resource Manager**
- Resource Manager
- Resource Manager
- **GRM Data**
- **GRM Data**
- **GRM Data**

**seL4 Address Spaces (VSpaces)**

- Very thin wrapper around hardware page tables
  - Architecture-dependent
  - ARM and (32-bit) x86 are very similar
- Page directories (PDs) map page tables, page tables (PTs) map pages
- A VSpace is represented by a PD object:
  - Creating a PD (by Retype) creates the VSpace
  - Deleting the PD deletes the VSpace

**Memory Management Mechanics: Retype**

- Retype (Untyped, $2^1$)
- Retype (Frame, $2^2$)
- Retype (CNode, $2^m$, $2^n$)
- Retype (TCB, $2^n$)
- Mint(r)
- Revoke()
### Address Space Operations

- Each mapping has:
  - `virtual_address`, `phys_address`, `address_space` and `frame_cap`
  - `address_space` struct identifies the level 1 page directory cap
  - you need to keep track of (`frame_cap`, `PD_cap`, `v_adr`, `p_adr`)

```c
seL4_Word frame_addr = ut_alloc(seL4_PageBits);
err = cspace_ut_retype_addr(frame_addr, seL4_ARM_Page,
  seL4_ARM_PageBits, cur_cspace, &frame_cap);
map_page(frame_cap, pd_cap, OxA0000000, seL4_AllRights,
  seL4_ARM_Default_VMAttributes);
bzero((void *)0xA0000000, PAGESIZE);
```

### Mapping Same Frame Twice: Shared Memory

- Each mapping requires its own frame cap even for the same frame

```c
seL4_CPtr new_frame_cap = cspace_copy_cap(cur_cspace,
  existing_frame_cap, seL4_AllRights,
  seL4_ARM_Default_VMAttributes);
map_page(new_frame_cap, pd_cap, OxA0000000, seL4_AllRights,
  seL4_ARM_Default_VMAttributes);
bzero((void *)0xA0000000, PAGESIZE);
```

### Memory Management Caveats

- The object manager handles allocation for you
- However, it is very simplistic, you need to understand how it works
- Simple rule (it’s buddy-based):
  - Freeing an object of size \( n \): you can allocate new objects \( \leq \) size \( n \)
  - Freeing 2 objects of size \( n \) does not mean that you can allocate an object of size \( 2n \).

<table>
<thead>
<tr>
<th>Object</th>
<th>size on ARM (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame</td>
<td>2^12</td>
</tr>
<tr>
<td>Page directory</td>
<td>2^14</td>
</tr>
<tr>
<td>Endpoint</td>
<td>2^4</td>
</tr>
<tr>
<td>Cslot</td>
<td>2^4</td>
</tr>
<tr>
<td>TCB</td>
<td>2^9</td>
</tr>
<tr>
<td>Page table</td>
<td>2^10</td>
</tr>
</tbody>
</table>

- All kernel objects must be size aligned!

### Memory Management Caveats

- Objects are allocated by `Retype()` of Untyped memory by seL4 kernel
  - `ut_alloc` and `ut_free()` manage user-level’s view of Untyped allocation.
  - Major pain if kernel and user’s view diverge
  - TIP: Keep objects address and CPtr together.
    - Be careful with allocations!
    - Don’t try to allocate all of physical memory as frames, as you need more memory for TCBs, endpoints etc.
    - Your frametable will eventually integrate with `ut_alloc` to manage the 4K untyped size.
Threads

- Threads are represented by TCB objects
- They have a number of attributes (recorded in TCB):
  - VSpace: a virtual address space
    - page directory reference
    - multiple threads can belong to the same VSpace
  - CSpace: capability storage
    - CNode reference (CSpace root) plus a few other bits
  - Fault endpoint
    - Kernel sends message to this EP if the thread throws an exception
  - IPC buffer (backing storage for virtual registers)
  - stack pointer (SP), instruction pointer (IP), user-level registers
  - Scheduling priority
  - Time slice length (presently a system-wide constant)

- These must be explicitly managed
  - … we provide an example you can modify

Creating a thread

- Obtain a TCB object
- Set attributes: Configure()
  - associate with VSpace, CSpace, fault EP, prio, define IPC buffer
- Set SP, IP (and optionally other registers): WriteRegisters()
  - this results in a completely initialised thread
  - will be able to run if resume_target is set in call, else still inactive
- Activated (made schedulable): Resume()

Creating a Thread in Own AS and cspace_t

```c
static char stack[100];
int thread_fct() {
    while(1);
    return 0;
}
/* Allocate and map new frame for IPC buffer as before */
seL4_Word tcb_addr = ut_alloc(seL4_TCBBits);
err = cspace_ut_retype_addr(tcb_addr, seL4_TCBObject, seL4_TCBBits,
                          cur_cspace, &tcb_cap);
err = seL4_TCB_Configure(tcb_cap, FAULT_EP_CAP, PRIORITY,
                          curspace->root_cnode, seL4NilData,
                          seL4_CapInitThreadPD, seL4_NilData,
                          PROCESS_IPC_BUFFER, ipc_buffer_cap);
seL4_UserContext context = { .pc = &thread, .sp = &stack};
seL4_TCB_WriteRegisters(tcb_cap, 1, 0, 2, &context);
```

If you use threads, write a library to create and destroy them.

Threads and Stacks

- Stacks are completely user-managed, kernel doesn’t care!
  - Kernel only preserves SP, IP on context switch
- Stack location, allocation, size must be managed by userland
- Beware of stack overflow!
  - Easy to grow stack into other data
  - Pain to debug!
  - Take special care with automatic arrays!
Creating a Thread in New AS and \texttt{cspace\_t}

```c
/* Allocate, retype and map new frame for IPC buffer as before
* Allocate and map stack???
* Allocate and retype a seL4\_ARM\_PageDirectoryObject of size seL4\_PageDirBits
* Mint a new badged cap to the syscall endpoint
*/
cspace_t * new_cspace = ut_alloc(seL4\_TCBBits);
char * elf_base = cpio_get_file(_cpio_archive, "test")->p_base;
err = elf_load(new_pagedirectory_cap, elf_base);
unsigned int entry = elf_getEntryPoint(elf_base);
err = seL4\_TCB\_Configure(tcb_cap, FAULT_EP_CAP, PRIORITY,
new_cspace->root_cnode, seL4NilData,
new_pagedirectory_cap, seL4\_NilData,
PROCESS\_IPC\_BUFFER, ipc_buffer_cap);

seL4\_UserContext context = {.pc = entry, .sp = &stack};
seL4\_TCB\_WriteRegisters(tcb_cap, 1, 0, 2, &context);
```

seL4 Scheduling

- Presently, seL4 uses 256 hard priorities (0–255)
  - Priorities are strictly observed
  - The scheduler will always pick the highest-prio runnable thread
  - Round-robin scheduling within prio level
- Aim is real-time performance, not fairness
  - Kernel itself will never change the prio of a thread
  - Achieving fairness (if desired) is the job of user-level servers

Exception Handling

- A thread can trigger different kinds of exceptions:
  - invalid syscall
    - may require instruction emulation or result from virtualization
  - capability fault
    - cap lookup failed or operation is invalid on cap
  - page fault
    - attempt to access unmapped memory
    - may have to grow stack, grow heap, load dynamic library, …
  - architecture-defined exception
    - divide by zero, unaligned access, …
- Results in kernel sending message to fault endpoint
  - exception protocol defines state info that is sent in message
- Replying to this message restarts the thread
  - endless loop if you don’t remove the cause for the fault first!

Exception Handling

- Exception triggered. Kernel fakes message from thread to handler
  - Handler performs appropriate action (e.g. map page).
- Kernel intercepts message and restarts thread
  - Handler replies to restart thread
Interrupt Management

- sel4 models IRQs as messages sent to an AEP
  - Interrupt handler has Receive cap on that AEP
- 2 special objects used for managing and acknowledging interrupts:
  - Single IRQControl object
    - single IRQControl cap provided by kernel to initial VSpace
    - only purpose is to create IRQHandler caps
  - Per-IRQ-source IRQHandler object
    - interrupt association and dissociation
    - interrupt acknowledgment

Interrupt Handling

- IRQHandler cap allows driver to bind AEP to interrupt
- Afterwards:
  - AEP is used to receive interrupt
  - IRQHandler is used to acknowledge interrupt

Device Drivers

- Drivers do three things:
  - Handle interrupts (already explained)
  - Communicate with rest of OS (IPC + shared memory)
  - Access device registers
- Device register access
  - Devices are memory-mapped on ARM
  - Have to find frame cap from bootinfo structure
  - Map the appropriate page in the driver’s VSpace

```
device_vaddr = map_device(0xA0000000, (1 << seL4_PageBits));
```

Magic device register access

Project Platform: i.MX6 Sabre Lite