



Microkernel-based Operating Systems —Introduction—

Jan Bierbaum, Carsten Weinhold

Dresden, October 15th 2024

- Provide deeper understanding of OS mechanisms
- Illustrate alternative design concepts
- Promote OS research at TU Dresden
- Make you all enthusiastic about OS development in general and microkernels in particular

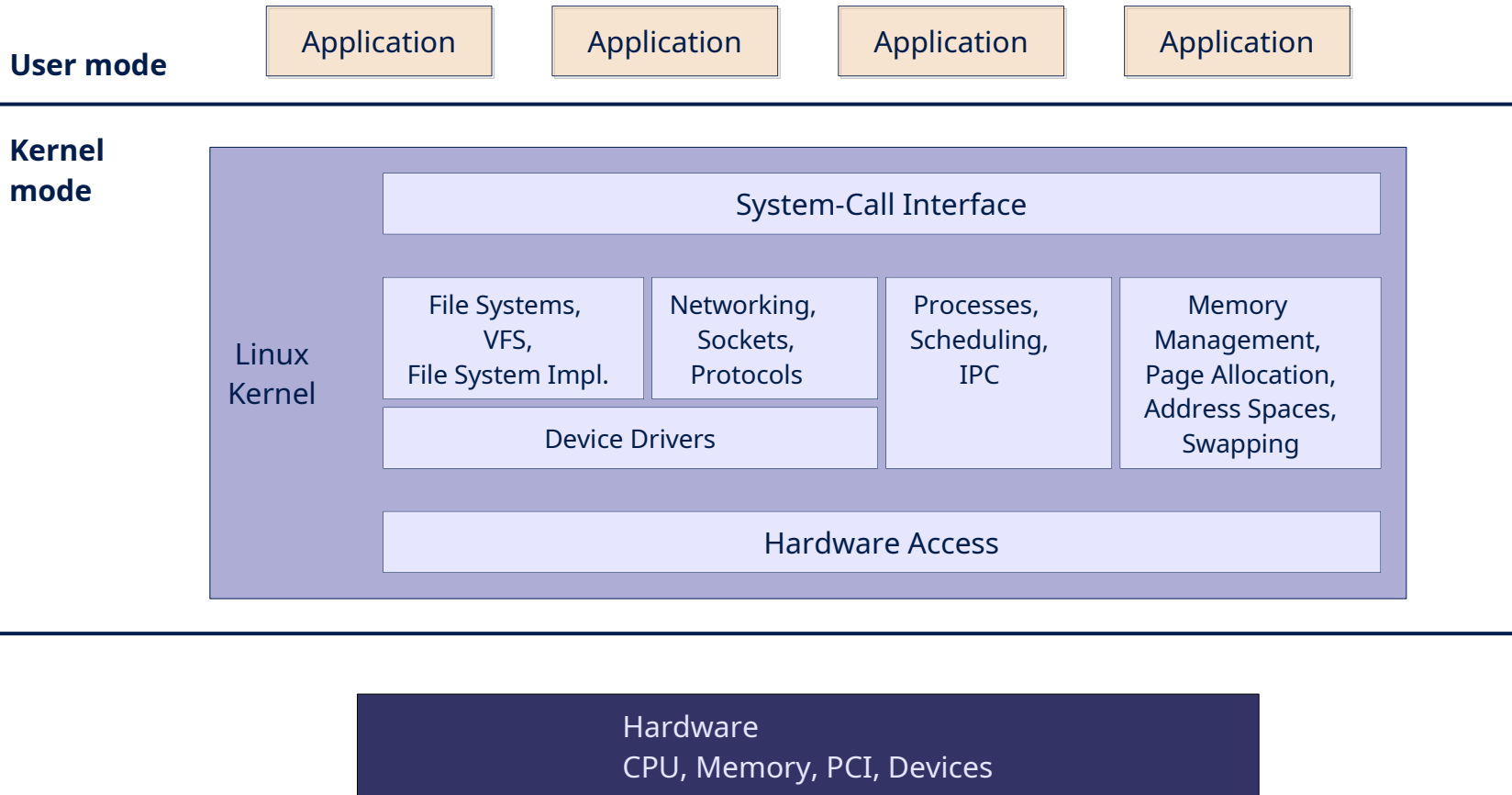
- Slides: <https://tud.de/inf/os> → Studies → Lectures → MOS
- Mailing list (announcement & updates): <https://os.inf.tu-dresden.de/mailman/listinfo/mos2024>
- Matrix: <https://matrix.to/#/#inf-os-mos:tu-dresden.de>
- Schedule:
 - Lecture: every Tuesday, 16:40, APB/E023
 - Exercises: (roughly) bi-weekly, Tuesday, 14:50
 - (mostly) hybrid via BBB
 - **Switch slots of lecture and exercise?**

- Practical exercises in the computer lab
 - Might happen at slightly different times
 - First one on *Nov 4th at 16:40*
- Paper reading exercises in APB/E008
 - Read a paper **beforehand**
 - Sum it up and prepare questions/observations
 - Actively participate in the discussion
 - First one *next week (Oct 22nd)*
- Exercises may also take place online/hybrid
- Announced on website and mailing list

- Complex lab “Microkernel-based Operating Systems” in parallel to this lecture
- Build several components of an MOS
- (roughly) bi-weekly, Tuesday, 14:50, APB E008
- Hybrid via BBB
- Mainly programming on your own
- Coordination via dedicated mailing list, check the [complex lab website](#)

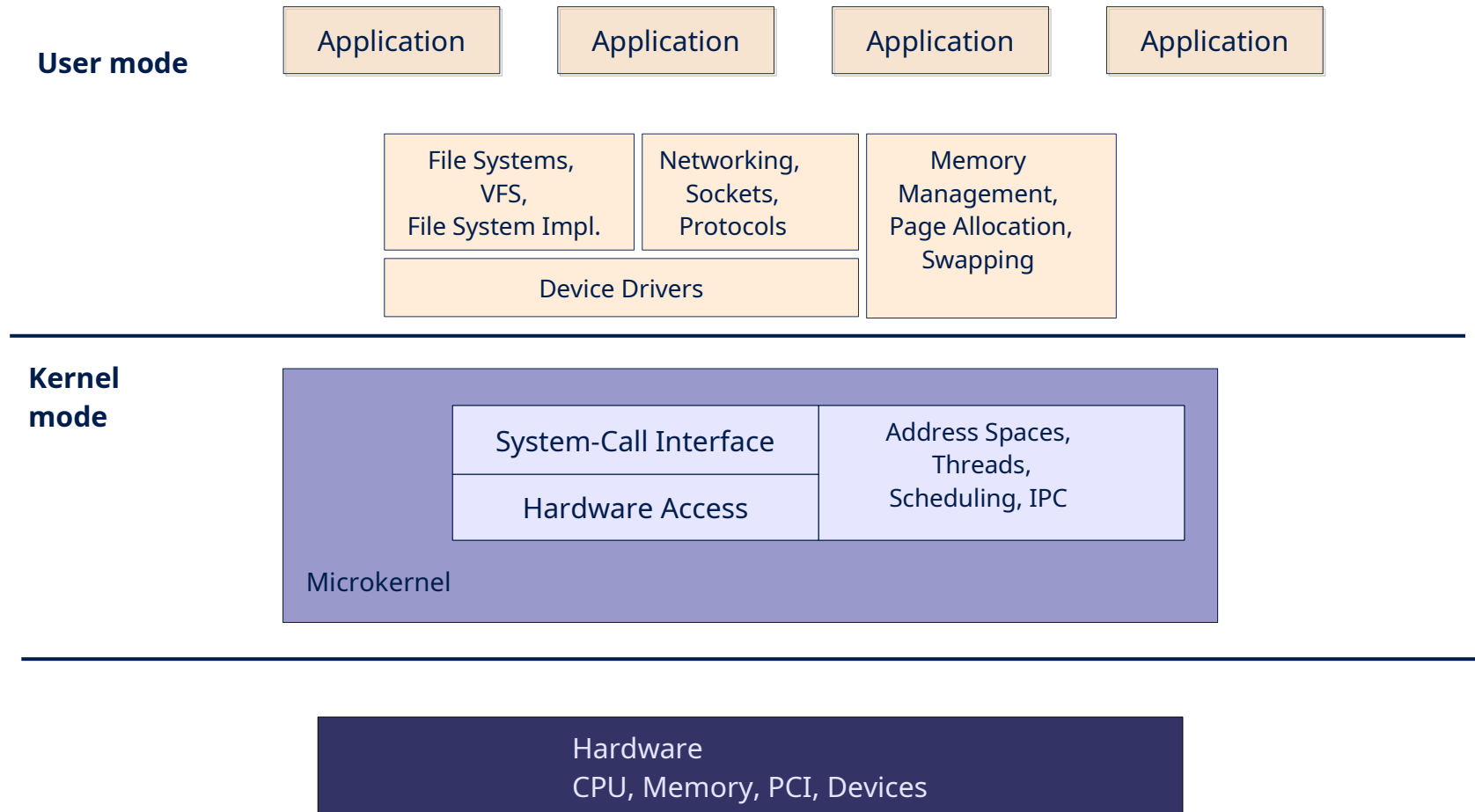
Monoliths vs. Microkernels

- Manage the available resources
 - Hardware (CPU, memory, storage, network, ...)
 - Software (file systems, networking stack, ...)
- Provide easier-to-use interface to access resources
 - Unix: read/write data from/to sockets instead of fiddling with TCP/IP packets on your own
- Perform privileged or HW-specific operations
 - x86: ring 0 vs. ring 3
 - Device drivers
- Provide separation and means for collaboration
 - Isolate users / processes from each other
 - Allow cooperation if needed (e.g., sending messages between processes)



- Security issues
 - All components in privileged mode
 - Direct access to all kernel-level data
 - Module loading → easy living for rootkits
- Resilience issues
 - Faulty drivers can crash the whole system
 - 75% of today's OS kernels are drivers
- Software-level issues
 - Complexity is hard to manage
 - Custom OS for hardware with scarce resources?

- Minimal OS kernel
 - Less error prone (less code → fewer errors)
 - Small *Trusted Computing Base*
 - Suitable for verification
- System services in user-level *servers*
 - Flexible and extensible
- Protection between individual components
 - More resilient: crashing component does not (necessarily...) crash the whole system
 - More secure: inter-component protection

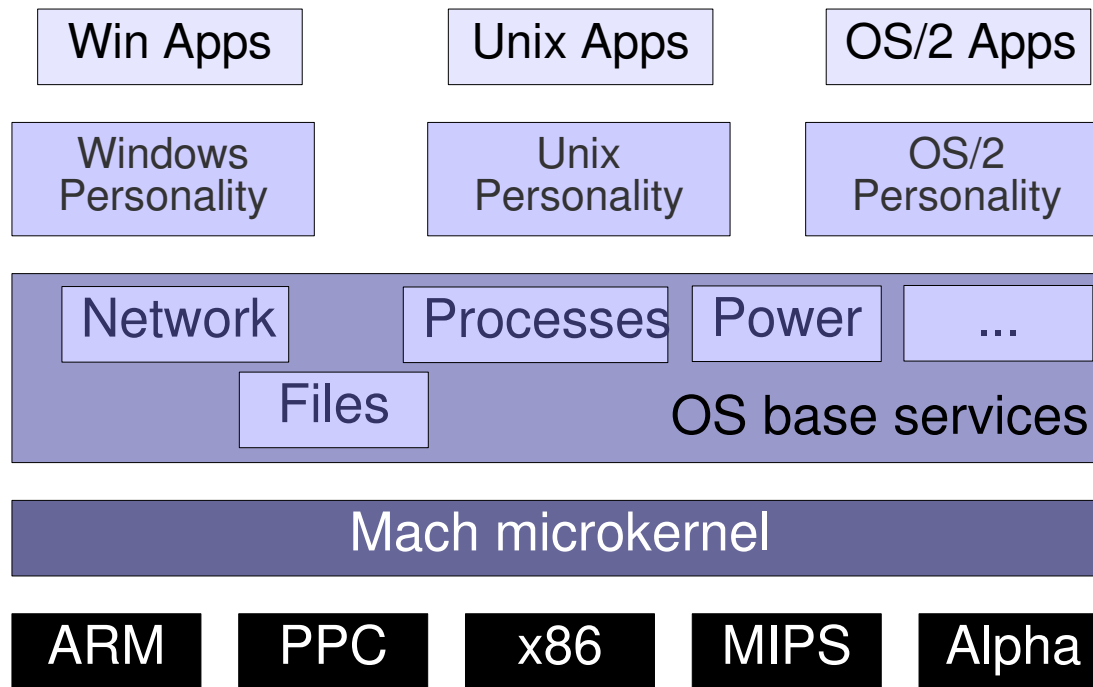


- OS personalities
- Customisability
 - Servers may be configured to suit the target system (small embedded systems, desktop PCs, SMP systems, ...)
 - Remove unnecessary servers
- Enforce reasonable system design
 - Well-defined interfaces between components
 - Access to components only via these interfaces
 - Improved maintainability

- Developed at CMU, 1985 – 1994
 - Rick Rashid (former head of MS Research)
 - Avie Tevanian (former Apple CTO)
 - Brian Bershad (professor @ U. of Washington)
 - ...
- Foundation for several real systems
 - Single Server Unix (BSD4.3 on Mach)
 - MkLinux (OSF, Apple)
 - IBM Workplace OS
 - NeXT OS → Mac OS X

- Simple, extensible *communication kernel*
 - “Everything is a pipe.”
 - *Ports* as secure communication channels
- Multiprocessor support
- Message passing by mapping
- Multi-server OS personality
- POSIX compatibility
- Shortcomings
 - Performance
 - Drivers in the kernel

- Main goals:
 - Multiple OS personalities
 - Support for multiple HW architectures



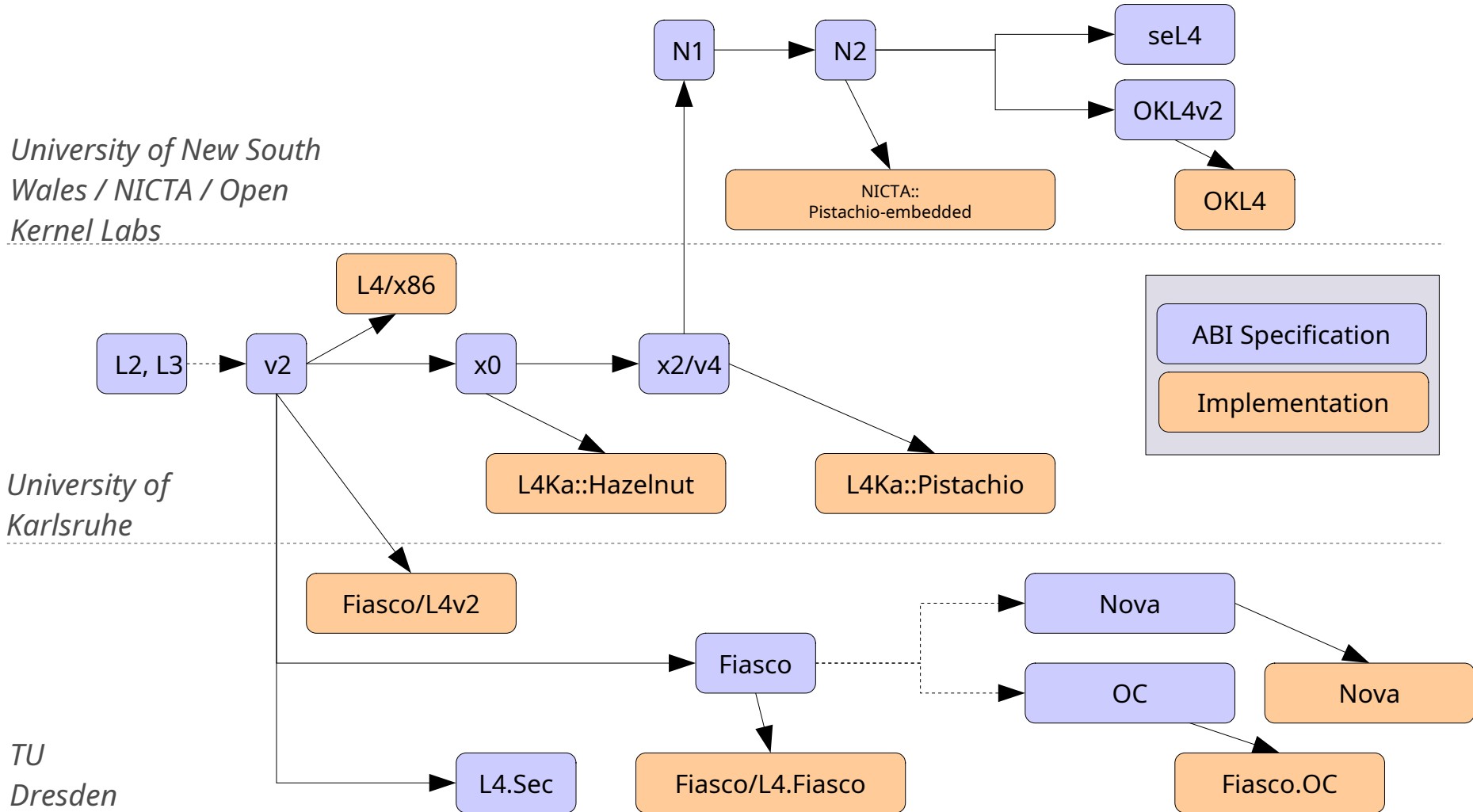
- Never finished ... but almost 2 billion US-\$ spent
- Causes of failure:
 - Underestimated difficulties in creating OS personalities
 - Management errors: divisions forced to adopt new system without having a system
 - “Second System Effect”: too many fancy features
 - Too slow
- Conclusion: Microkernel worked, but system atop the microkernel did not

- OS personalities did not work
- Flexibility ... but monolithic kernels became flexible, too (e.g., Linux kernel modules)
- Better design ... but monolithic kernels also improved (restricted symbol access, layered architectures)
- Maintainability ... still very complex
- Performance matters a lot

- Subsystem protection / isolation
- Code size (generated using David A. Wheeler's "SLOCCount")
 - Microkernel-based OS
 - Fiasco kernel: ~34,000 LoC
 - "HelloWorld" (+ boot loader + root task): ~10,000 LoC
 - Linux kernel (3.0.4., x86 architecture):
 - Kernel: ~2.5 million LoC
 - + Drivers: ~5.4 million LoC
- Customisability
 - Tailored memory management / scheduling / ... algorithms
 - Adaptable to embedded / real-time / secure / ... systems

- We need fast and efficient kernels
 - Covered in the “Microkernel Construction” lecture in the summer term
- We need fast and efficient OS services
 - Memory and resource management
 - Synchronisation
 - Device Drivers
 - File systems
 - Communication interfaces
 - Subject of this lecture

- Minix @ FU Amsterdam (Andrew Tanenbaum)
- Singularity @ MS Research
- EROS/CoyotOS @ Johns Hopkins University
- The L4 Microkernel Family
 - Originally developed by Jochen Liedtke at IBM and GMD
 - 2nd-generation microkernel
 - Several kernel ABI versions



- Jochen Liedtke:
“A microkernel does no real work.”
 - Kernel only provides inevitable mechanisms.
 - Kernel does not enforce policies.
- But what **is** inevitable?
 - Abstractions
 - Threads
 - Address spaces (tasks)
 - Mechanisms
 - Communication
 - Resource mapping
 - (Scheduling)

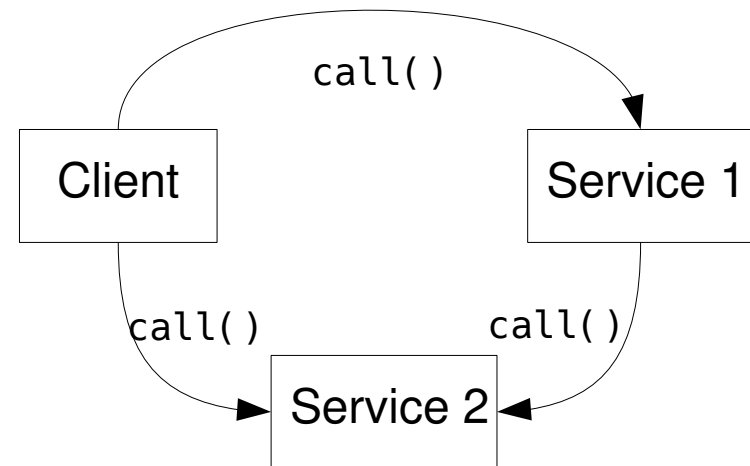
Taking a closer look at L4

Case study: **L4/Fiasco.OC**

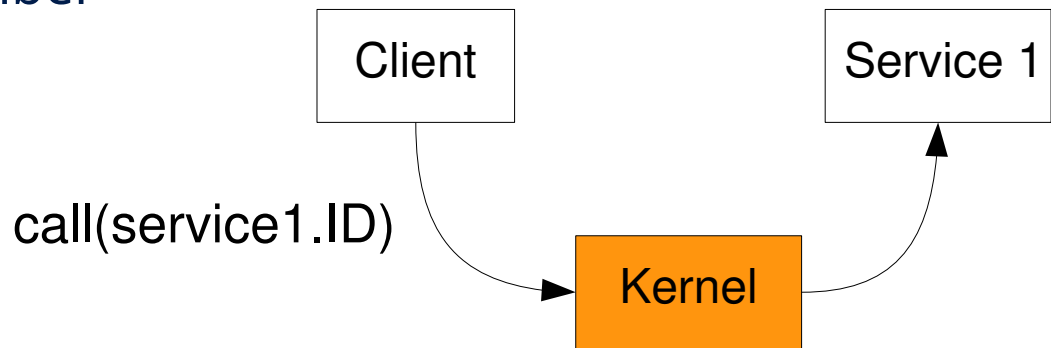
- “Everything is an object”
 - Task **Address spaces**
 - Thread **Activities, scheduling**
 - IPC Gate **Communication, resource mapping**
 - IRQ **Communication**
 - Factory Create other objects, enforce
 resource quotas
- One system call: **invoke_object()**
 - Parameters passed in UTCB
 - Types of parameters depend on type of object

- Kernel-provided objects
 - Threads
 - Tasks
 - IRQs
 - ...
- Generic communication object: IPC gate
 - Send message from sender to receiver
 - Allows to implement **new objects** in **user-level** applications

- Build everything above kernel using user-level objects that provide a service
 - Networking stack
 - File system
 - ...
- Kernel provides
 - Object creation/management
 - Object interaction: Inter-Process Communication (IPC)



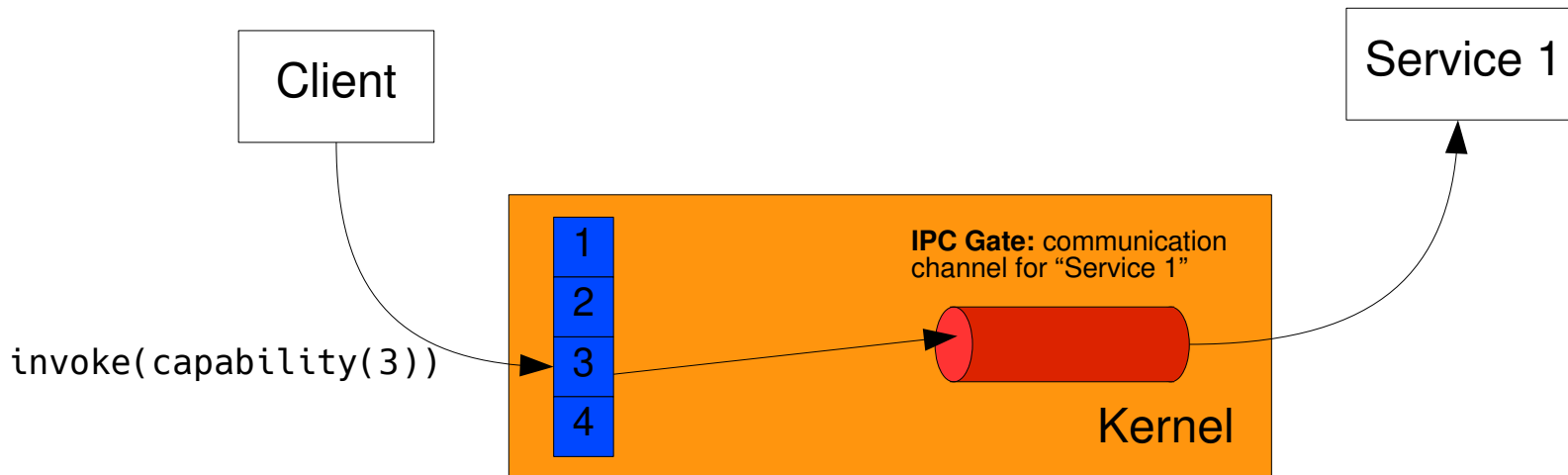
- To call an object, we need an address:
 - Telephone number
 - Postal address
 - IP address
 - ...



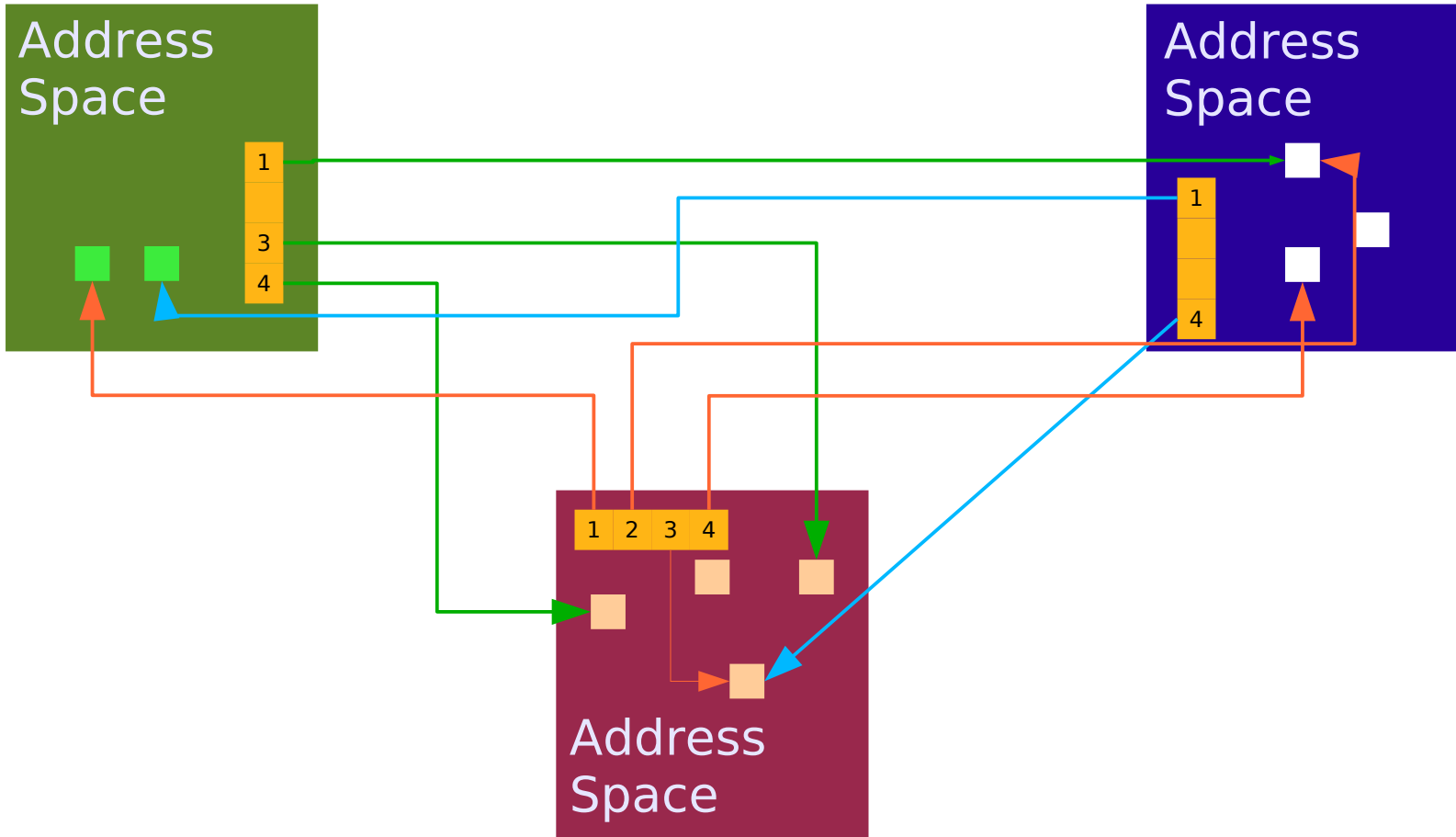
- Simple idea, isn't it?
- ID is wrong? Kernel returns ENOTEXIST
- But not so fast! This scheme is insecure:
 - Client could simply “guess” IDs brute-force
 - (Non-)Existence can be used as a covert channel

- Global object IDs are
 - insecure (forgery, covert channels)
 - inconvenient (programmer needs to know about partitioning in advance)
- Solution in Fiasco.OC
 - Task-local *capability space* as indirection
 - *Object capability* required to invoke object
 - Per-task name space
 - Maps names to object capabilities
 - Configured by task's creator

- Capability:
 - Reference to an object
 - Protected by the kernel
 - Kernel knows all capability-object mappings
 - Managed as a per-process capability table
 - User processes only use indices into this table

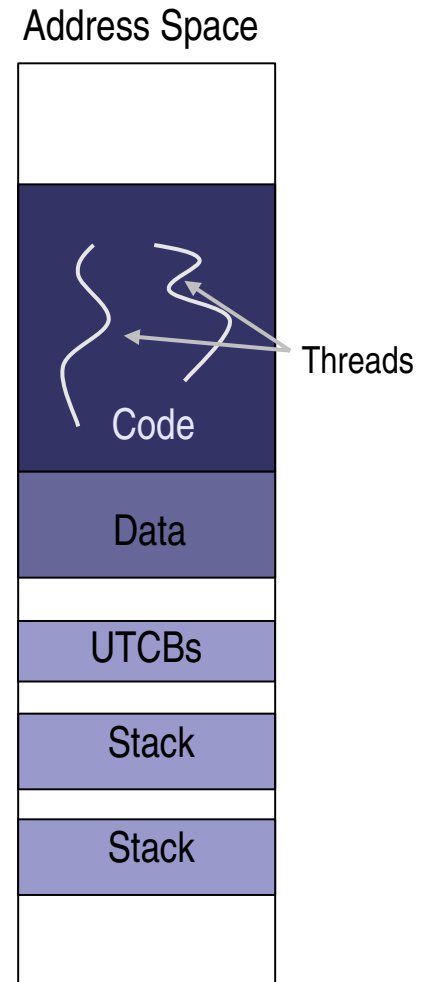


- Kernel object for communication: *IPC gate*
- Inter-process communication (IPC)
 - Between threads
 - Synchronous
- Sequence:
 - Sender writes message into its UTCB
 - Sender invokes IPC gate → blocks sender until receiver ready (i.e., waits for message)
 - Kernel copies message to receiver thread's UTCB
 - Both continue, knowing that message has been transferred/received

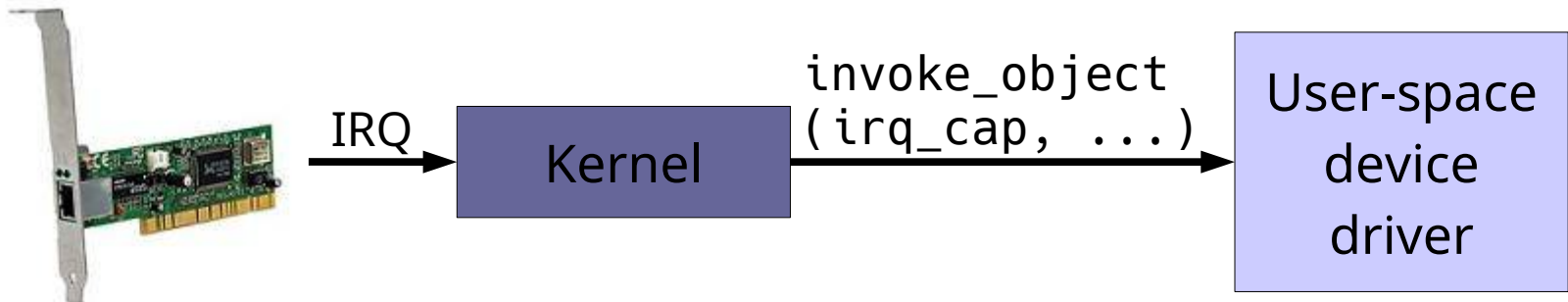


More L4 concepts

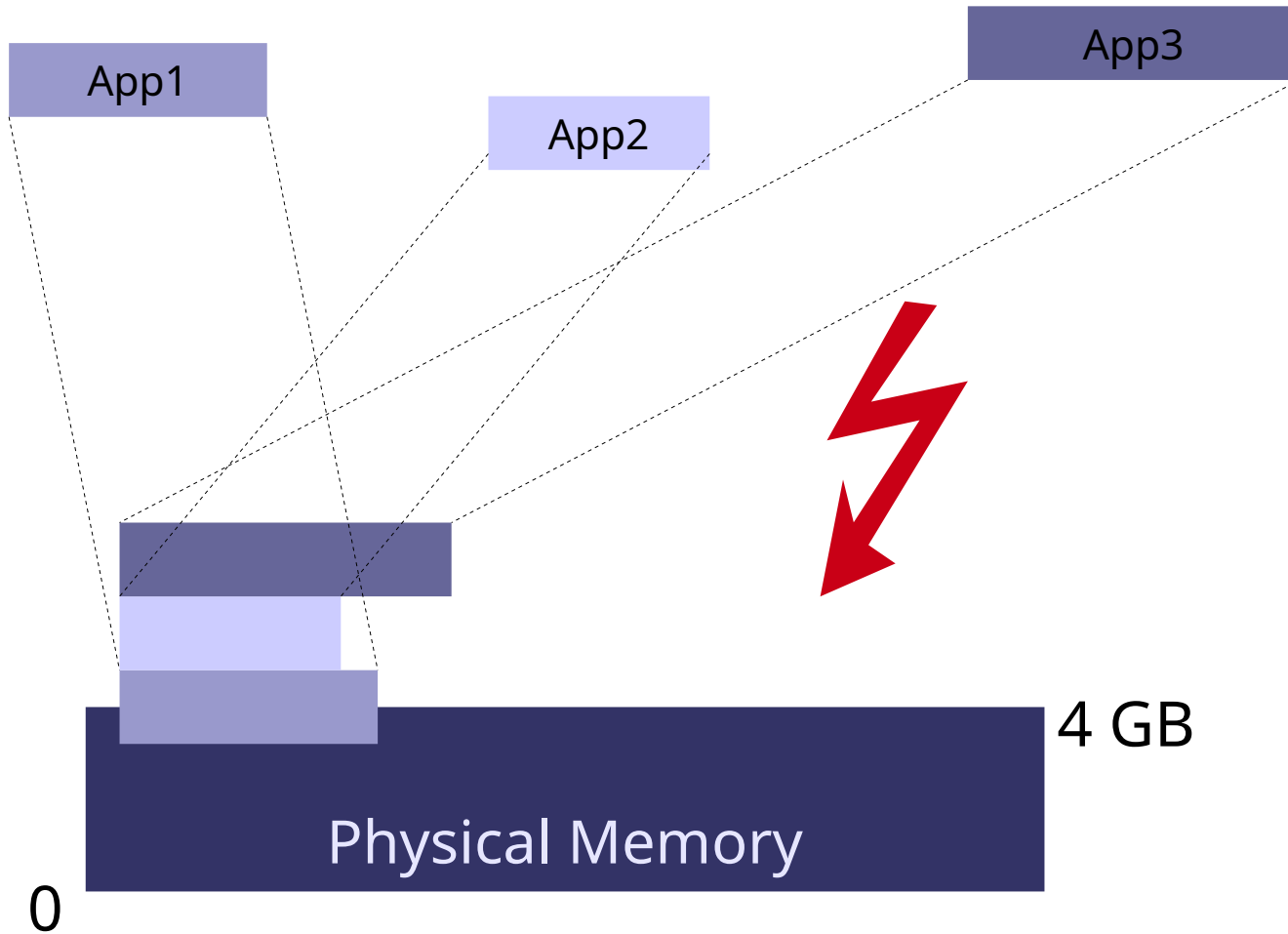
- Thread
 - Unit of execution
 - Implemented as kernel object
- Properties managed by the kernel
 - Instruction Pointer (EIP)
 - Stack (ESP)
 - Registers
 - User-level thread control block (UTCB)
- User-level applications need to
 - allocate stack memory
 - provide memory for application binary
 - find entry point
 - ...

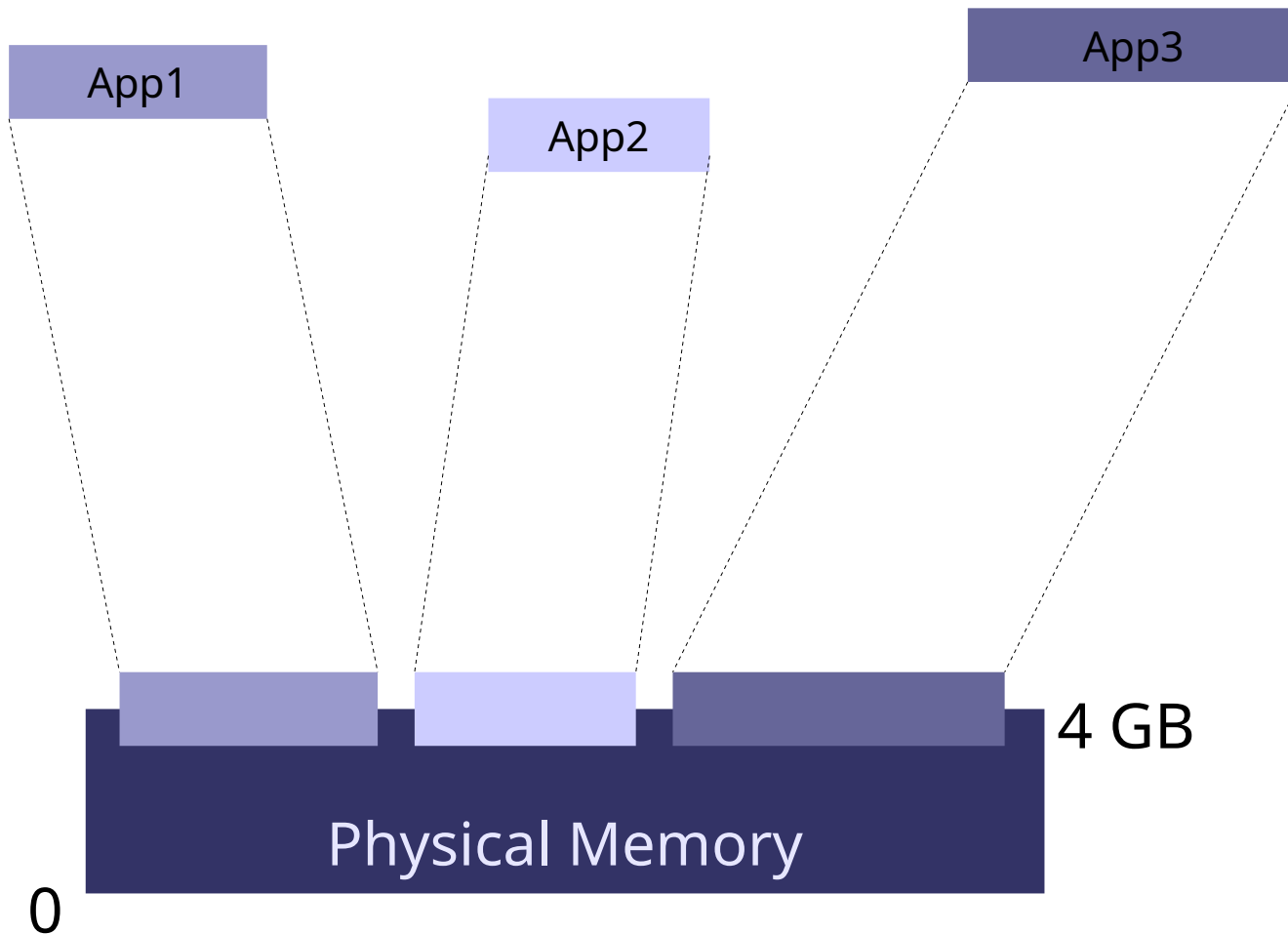


- Kernel object: IRQ
- Used for hardware and software interrupts
- Provides asynchronous signaling
 - `invoke_object(irq_cap, WAIT)`
 - `invoke_object(irq_cap, TRIGGER)`

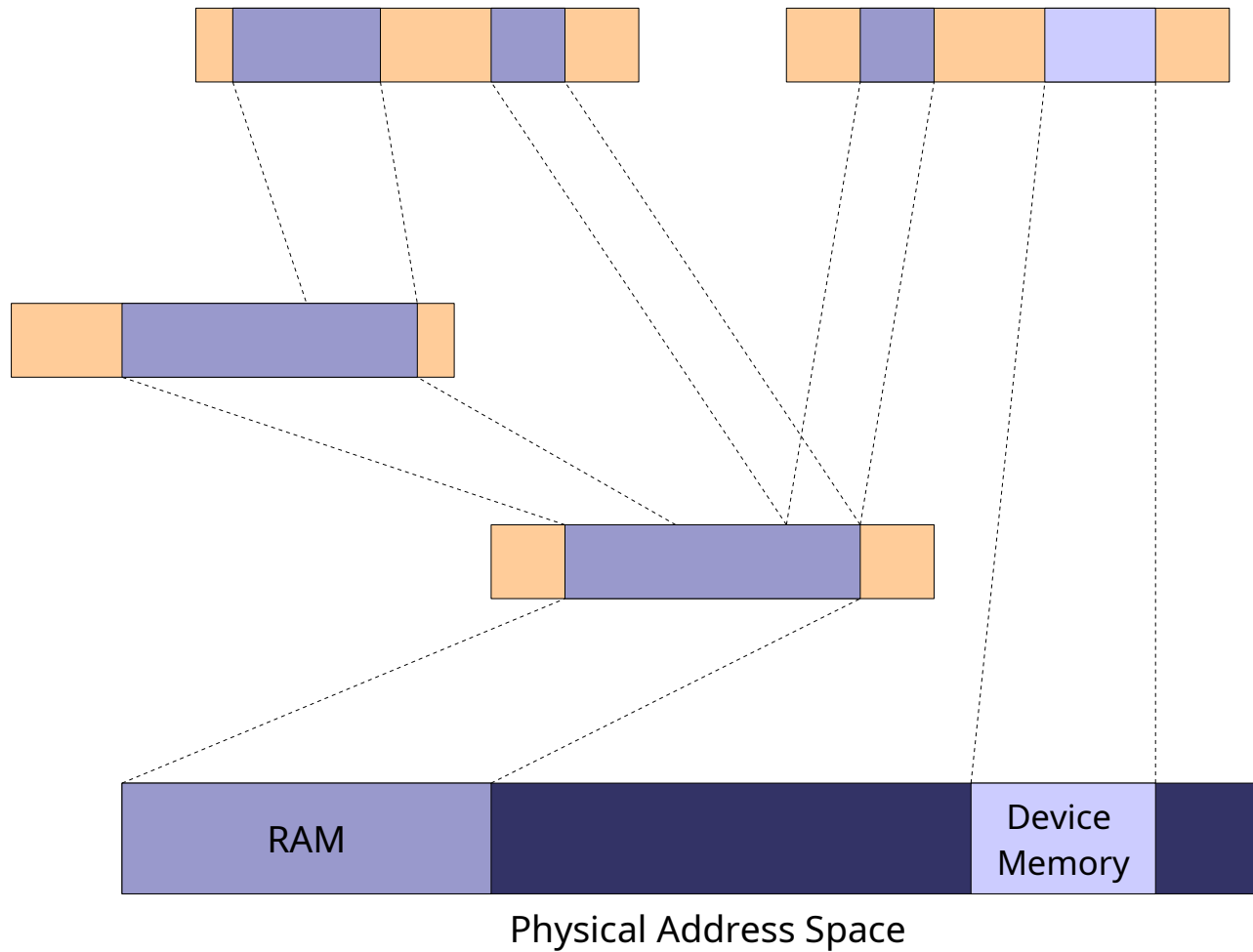


Challenge: Memory partitioning





L4: Recursive Address Spaces

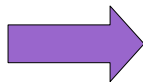


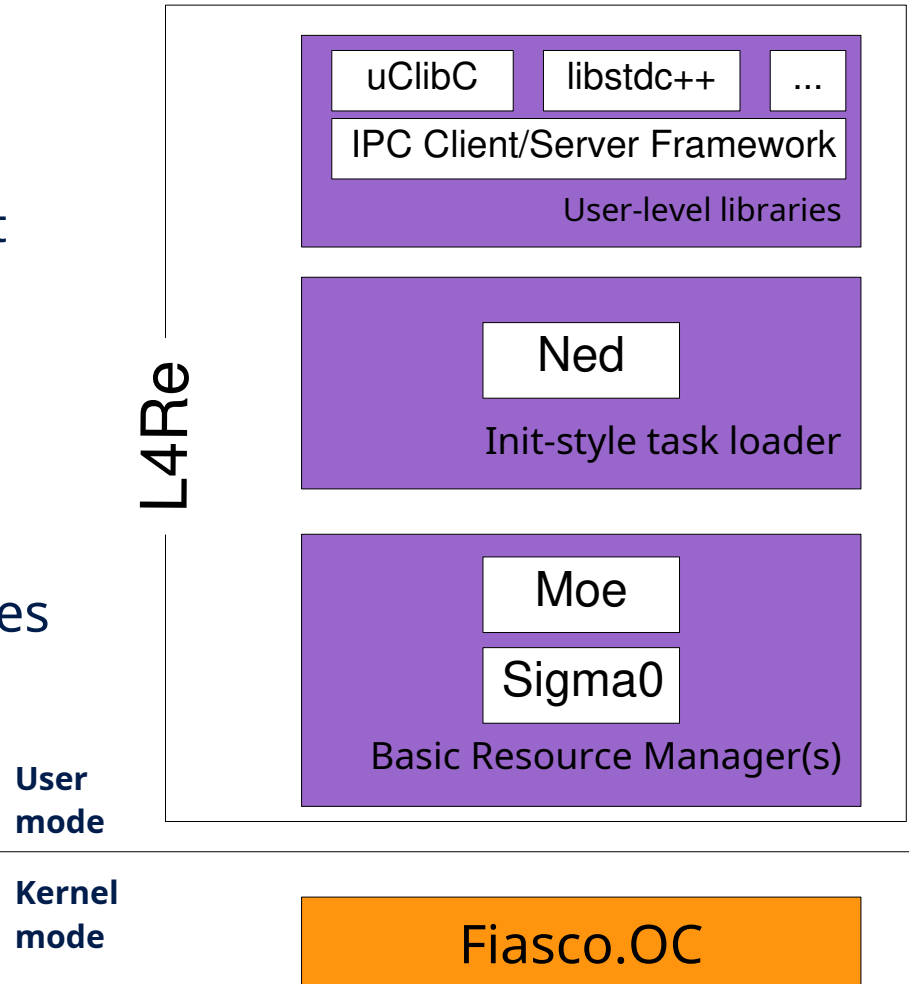
- If a thread has access to a capability, it can map this capability to another thread
- Mapping / not mapping of capabilities used to implement access control
- Abstraction for mapping: *flexpage*
 - Location and size of resource
 - Receiver's rights (read-only, mappable)
 - Type (memory, I/O, communication capability)

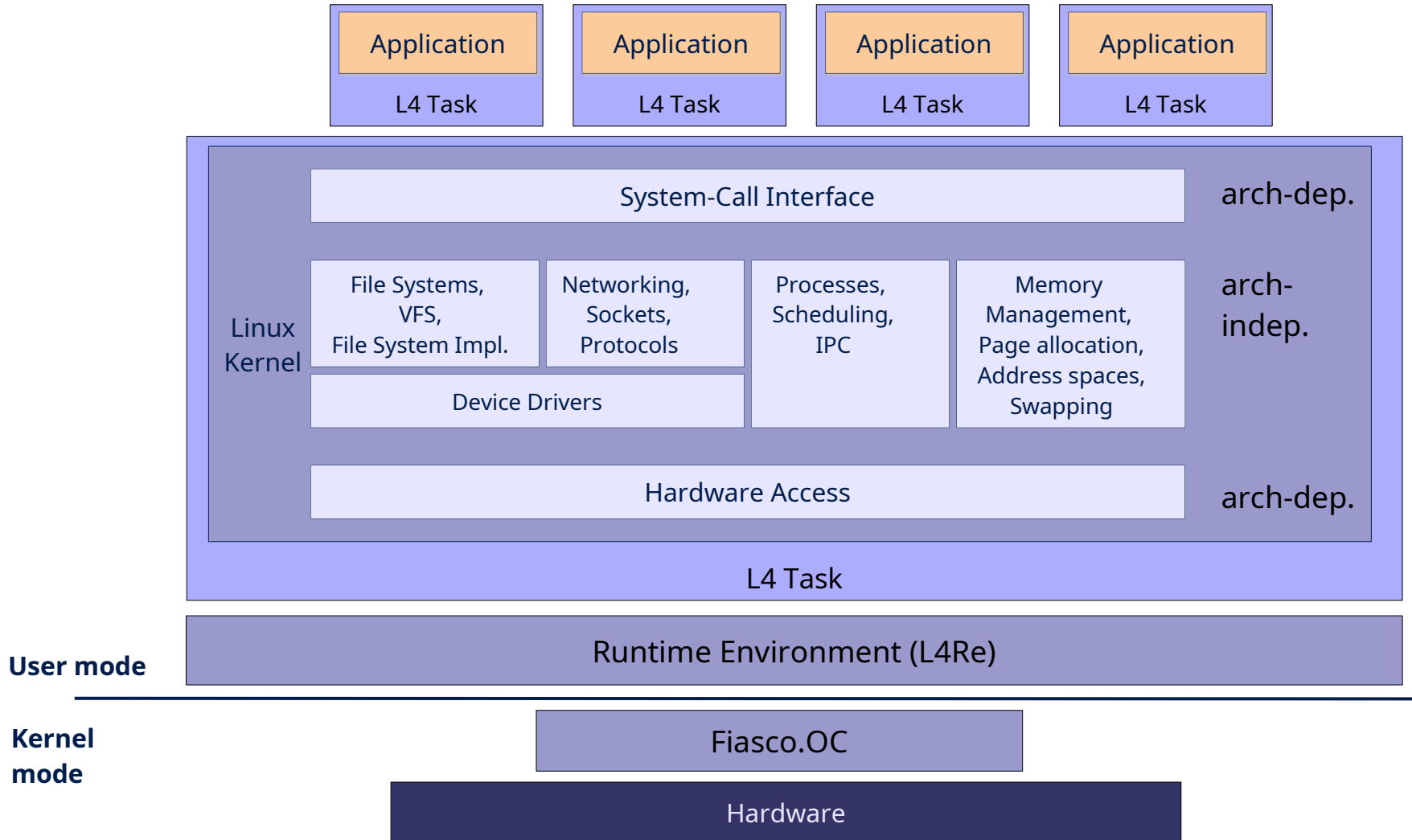
- Summary of object types
 - Task
 - Thread
 - IPC Gate
 - IRQ
 - Factory
- Each task gets initial set of capabilities for some of these objects at startup

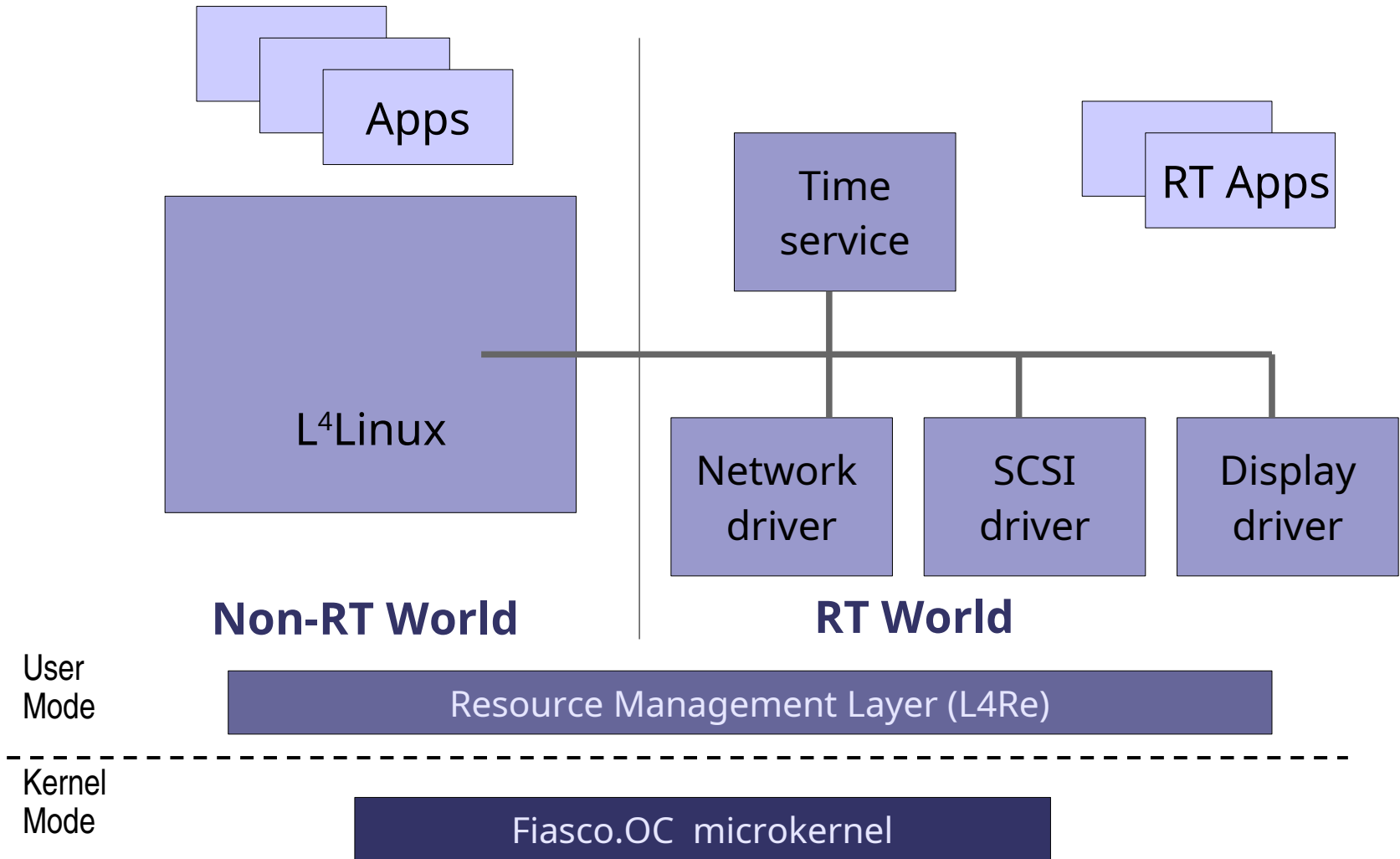
What can we build with this?

- Fiasco.OC is not a full operating system!
 - No device drivers (except UART + timer)
 - No file system / network stack / ...
- A microkernel-based OS needs to add these services as user-level components

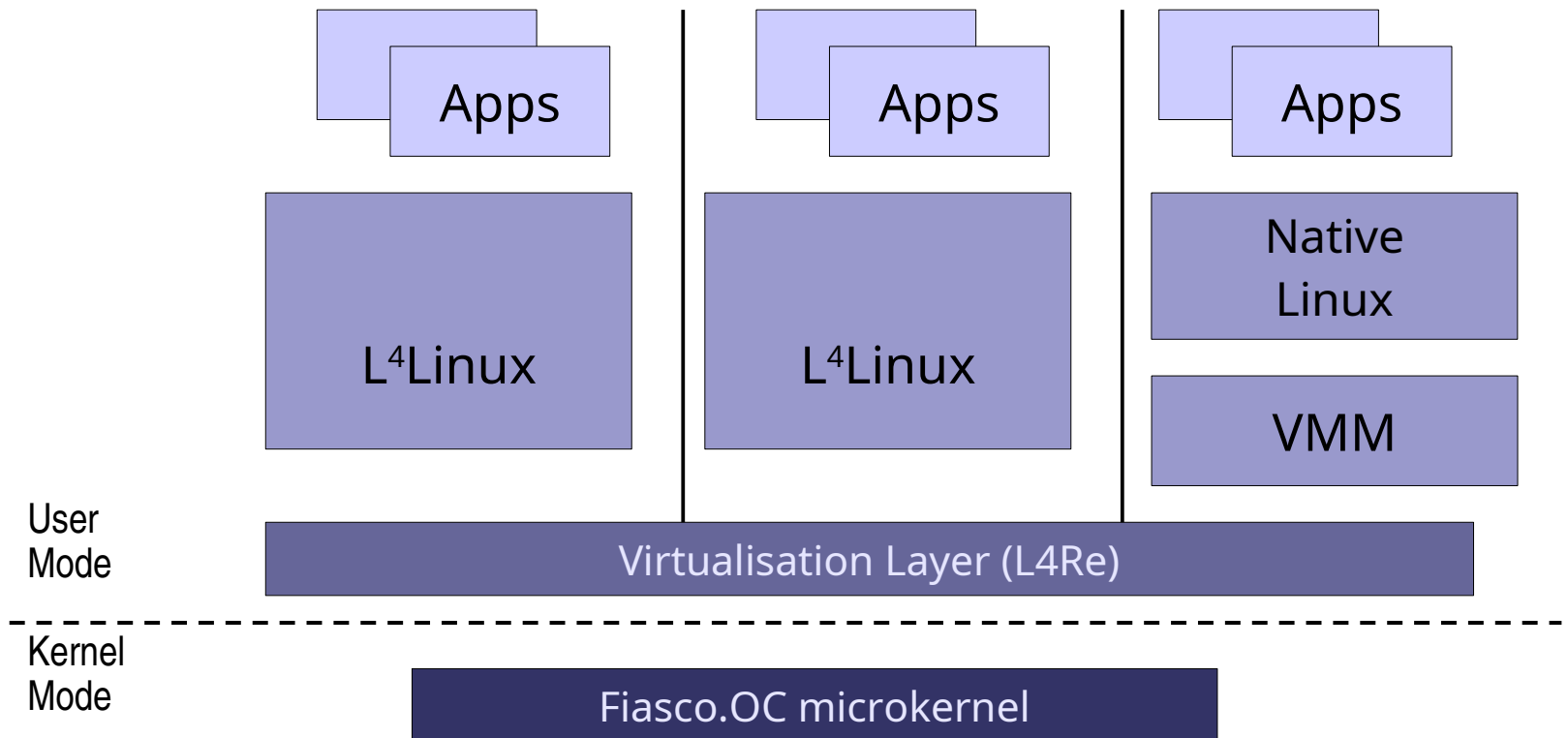

L4Re = L4 Runtime Environment



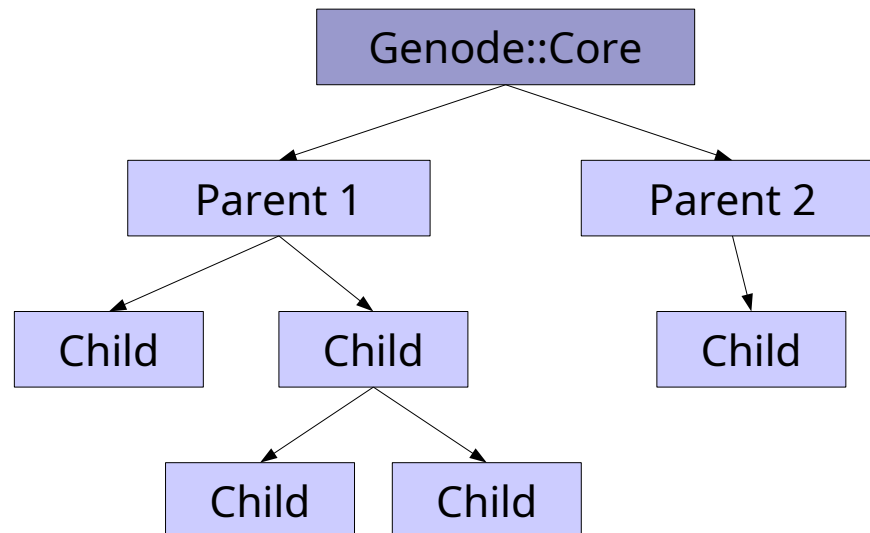




- Isolate not only processes, but also complete operating systems (compartments)
- “Server consolidation”



- Genode = C++-based OS framework developed here in Dresden
- Goal: hierarchical system that
 - supports resource partitioning
 - layers security policies on top of each other



What's to come?

- **Basic mechanisms and concepts**
 - Memory management
 - Tasks, Threads, Synchronisation
 - Communication
- **Building real systems**
 - What are resources and how to manage them?
 - How to build a secure system?
 - How to build a real-time system?
 - How to re-use existing code (Linux, standard system libraries, device drivers, ...)?
 - How to improve robustness and safety?

- Next lecture:
 - No lecture next week
 - In **two** weeks (Oct 29th): *“Inter-Process Communication (IPC)”*
- First exercise:
 - Next week (Oct 22nd): Per Brinch Hansen — *“The nucleus of a multiprogramming system”*
 - Link on the website
 - **Read the paper!**