



CS 410/510

Languages & Low-Level Programming

Mark P Jones
Portland State University

Spring 2016

Week 5: The L4 Microkernel

From ad-hoc to generic

- So far, we've been building bare-metal applications in an ad-hoc manner
- ... which would be reasonable in a custom embedded system
- but what if we want a more generic, reusable foundation for building and deploying computer systems?
- (also known as an "operating system" 😊)
- Let's take a look at L4 as an initial case study ...

1

2

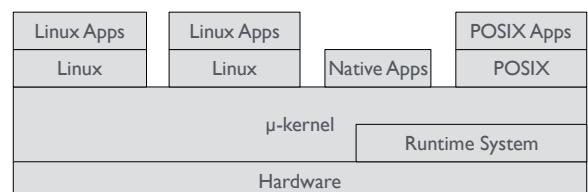
Why L4?

it came in a vision ...

3



But seriously: Context ...



In the "Programatica" project, we were looking to build a μ-kernel with very high assurance of separation between domains

6

Microkernel philosophy

- Minimize the amount of code that runs in kernel mode; implement other functionality in “user space servers”
 - Minimize the “Trusted Computing Base” (TCB)
- A microkernel must abstract physical memory, CPU (threads), and interrupts/exceptions
- A microkernel must also provide (efficient) mechanisms for communication and synchronization
- A microkernel should be “policy free”

Microkernel design: L4

- L4 is a “second generation” μ-kernel design, originally designed by Jochen Liedtke
- Designed to show that μ-kernel based systems are usable in practice with good performance
- Minimalist philosophy: If it can be implemented outside the kernel, it doesn’t belong inside

7

8

Why pick L4?

- L4 is industrially and technically relevant
 - Multiple working implementations (Pistachio, Fiasco, OKL4, etc...)
 - Multiple supported architectures (ia32, arm, powerpc, mips, sparc, ...)
 - Already used in a variety of domains, including real-time, security, virtual machines & monitors, etc...
 - Open Kernel Labs spin-off from NICTA & UNSW
 - Commercial use by Qualcomm and others ...

Why pick L4?

- L4 is industrially and technically relevant
- L4 is small enough to be tractable
 - Original implementation ~ 12K executable
 - Recent/portable/flexible implementations ~ 10-20 KLOC C++
 - Our original plans called for a POSIX implementation ... scary!

9

10

Why pick L4?

- L4 is industrially and technically relevant
- L4 is small enough to be tractable
- L4 is real enough to be interesting
 - For example, we can run multiple, separated instances of Linux (specifically: L4Linux, Wombat) on top of an L4 μ-kernel
 - Use somebody else’s POSIX layer rather than build our own!
 - Detailed specification documents are available

Why pick L4?

- L4 is industrially and technically relevant
- L4 is small enough to be tractable
- L4 is real enough to be interesting
- L4 is a good representative of the target domain and a good tool for exposing core research challenges
 - Threads, address spaces, IPC, preemption, interrupts, etc... are core μ-kernel concepts, regardless of API details
 - It should be possible to retarget to a different API or μ-kernel design

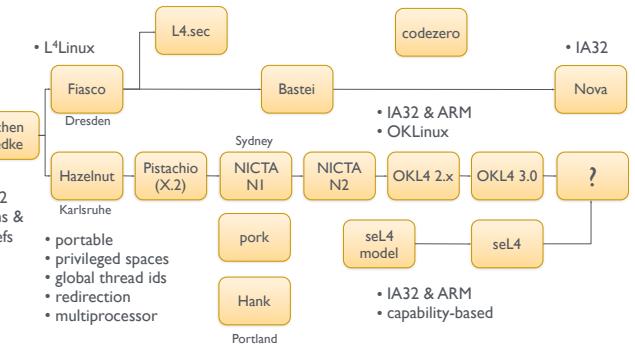
11

12

Why pick L4?

- L4 is industrially and technically relevant
- L4 is small enough to be tractable
- L4 is real enough to be interesting
- L4 is a good representative of the target domain and a good tool for exposing core research challenges
- L4 is “not invented here”
 - We’re not in the business of OS design and implementation
 - Leverage the insights and expertise of the OS community so that we can focus on our own research goals
 - A credibility boost, showing that our methods apply to other people’s problems (we can’t change the OS design to make our lives easier ...)

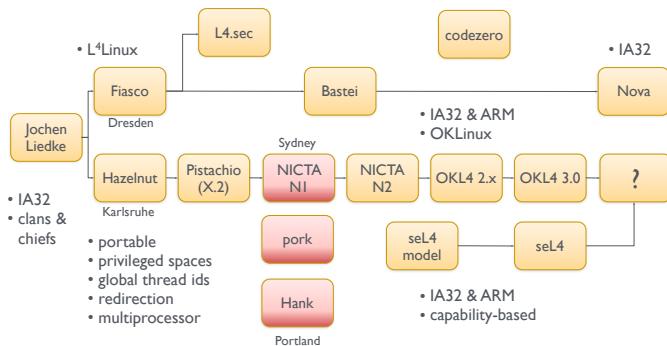
Evolution of L4



13

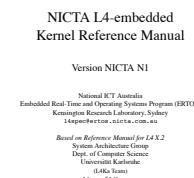
14

Evolution of L4 - Case Study I



NICTA NI

- For concreteness, this presentation will be based (mostly) on the NICTA NI version of the L4 spec
- Available in reference section of D2L course content
- (primary reference for pork)
- Lots of diagrams of bitdata and memory area structures
- ... implications for language design?

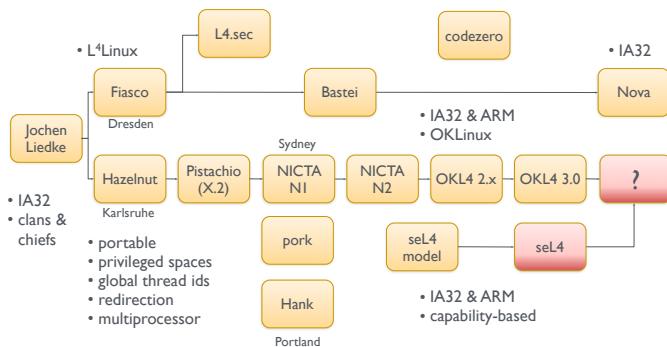


Document Revision 2
October 7, 2005

15

16

Evolution of L4 - Case Study 2

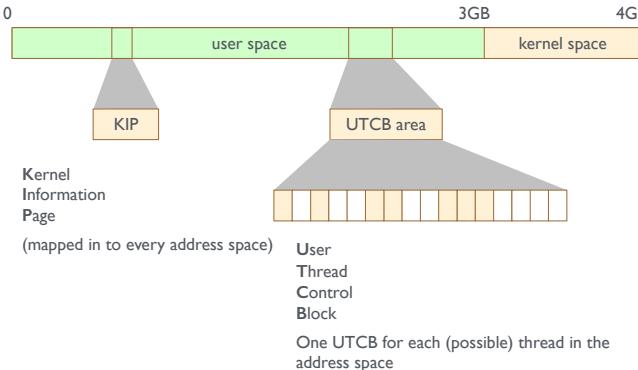


Address Space Layout

17

18

Userspace perspective



What's in the KIP?

- Information about the kernel version

~	SCHEDULE SC	THREADSWITCH SC	Reserved	+FO / +IEO
EXCHANGEREGISTERS SC	UNIMAP SC	LIBC SC	IBC SC	+FO / +HCO
MEMORYCONTROL jSC	API Version	version (8)	subversion (8)	~ (16)
ProcessorInfo		version	subversion	
ProcDescPtr		0x02		A0
KipAreaInfo		0x83	0x80	B0
		0x83	0x81	60
		0x84	rev	40
		0x85	rev	20
		0x86	rev	00
		0x04	rev	
	API Flags			
				~ (28/16/8)
cc		= 00 : little endian,		E0
		= 01 : big endian,		C0
ww		= 00 : 32 bit API,		A0
		= 01 : 64 bit API,		80
		~		+30 / +60
		~		+20 / +40
		~		+10 / +20
KernDescPtr	API Flags	API Version	0_0_(32)	+0
			X Y Z T	
-FC / +18		+8 / +10		
		+4 / +8		

19

20

What's in the KIP?

- Information about the kernel version
 - Information about the host system

ProcessorInfo	<code># (4)</code> <code>\sim (12/44)</code> <code>processor# = 1 (16)</code>	
*	The size of the area occupied by a single processor description is 2^k . Location of description fields for the first processor is denoted by <code>ProcDescP0</code> . Description fields for subsequent processors are located directly following the previous one.	
Processors	<code>Number of available system processors.</code>	<code>MemoryInfo</code>
PageInfo	<code>page-size mask (22/54)</code> <code>\sim (7)</code> <code>FWE#</code>	
<code>page-size mask</code>	If this bit = 10 of the page-size mask field (bit 10 of the entire word) is set to 1 hardware and kernel support pages of size 2^k . If this bit is 0 hardware and/or kernel do not support pages of size 2^k . Note that pages of size 2^k can be used, even if 2^k is not supported hardware page size. Information about supported hardware page sizes is only a performance hint.	<code>API Version</code> <code>0 (0/32)</code> <code>KC1208 4^t</code>

What's in the KIP?

- Information about the kernel version
 - Information about the host system
 - Information about address space layout

	SCHEDULE SC	THREADSWITCH SC	Reserved	
EXCHANGEREGISTERS SC	UNMAP SC	LIPC SC	IIPC SC	+FO / +IE0
MEMORYCONTROL_pSC	PROCESSORCONTROL_pSC	THREADCONTROL_pSC	SPACECONTROL_pSC	+EO / +ICO
ProcessEndInfo	PageInfo	ThreadInfo	ClockInfo	+DO / +IA0
ProcDespPtr	BootInfo		~	+CO / +180
KipAreaInfo	UtehInfo	VirtualRegInfo	~	+BO / +160
			~	+AO / +140
			~	+9O / +120
			~	+8O / +100
			~	+7O / +EO
			~	+6O / +CO
		MemoryInfo	~	+5O / +AO
			~	+4O / +80
			~	+3O / +60
			~	+2O / +40
			~	+1O / +20
			~	+0

UtehInfo

~ (10/42)	s (6)	a (6)	m (10)
-----------	-------	-------	--------

s The minimal area size for an address space's UTCB area is 2^s . The size of the UTCB area limits the total number of threads k to $2^m \leq 2^s$. A size of 0 indicates that the UTCB is not part of the user address space and cannot be controlled (see page 41).

m UTCB size multiplier.

a The UTCB location must be aligned to 2^a . The total size required for one UTCB is $2^a m$.

2

22

What's in the KIP?

- Information about the kernel version
 - Information about the host system
 - Information about address space layout
 - System call entry points
 - So how can a user process find the KIP address?

~	SCHEDULE SC	THREADSWITCH SC	Reserved
EXCHANGEREGISTERS SC	UNMAP SC	LIPC SC	IIPC SC
MEMORYCONTROL pSC	PROCESSORCONTROL pSC	THREADCONTROL pSC	SPACECONTROL pSC
ProcessorInfo	PageInfo	ThreadInfo	ClockInfo
ProcDescPtr	BoolInfo		~
KipAreaInfo	UtsInfo	malRegInfo	~
Offsets, relative to the start of the KIP to system call entry points			
KernelDescPtr	APIFlags	API Version	0(0)/32K/256/4
+C / +18	+8 / +10	+4 / +8	

Offsets, relative to the start of the KIP to system call entry points

How to find the KIP

- Option 1: Design protocol
 - User code assumes a predetermined KIP address
 - Option 2: “Slow system call” ... a “virtual” instruction
 - User code executes the illegal instruction **LOCK NOP**
 - This triggers an illegal opcode exception, which enters the kernel
 - The kernel checks for this exception, loads the kip address in to the context registers, and returns to user mode

	KernelInterface	
- EAX	lock:	EAX <i>base address</i>
- ECX	nop	ECX <i>API Version</i>
- EDX		EDX <i>API Flags</i>
- ESI		ESI <i>Kernel ID</i>
- EDI		EDI \equiv
- EBX		EBX \equiv
- EBP		EBP \equiv
- ESP		ESP \equiv

23

24

What are the gaps for?

\sim	SCHEDULE SC	THREADSWITCH SC	Reserved	+F0 / +1E0
EXCHANGEREGISTERS SC	UNMAP SC	LIPC SC	IPC SC	+E0 / +1C0
MEMORYCONTROL pSC	PROCESSORCONTROL pSC	THREADCONTROL pSC	SPACECONTROL pSC	+D0 / +1A0
ProcessorInfo	PageInfo	ThreadInfo	ClockInfo	+C0 / +180
ProcDescPtr	BootInfo		\sim	+B0 / +160
KipAreaInfo	UtblInfo	VirtualRegInfo	\sim	+A0 / +140
		\sim		+90 / +120
		\sim		+80 / +100
		\sim		+70 / +E0
		\sim		+60 / +C0
Kdebug.config1	Kdebug.config0	MemoryInfo	\sim	+50 / +A0
root server.high	root server.low	root server.IP	root server.SP	+40 / +80
σ_1 .high	σ_1 .low	σ_1 .IP	σ_1 .SP	+30 / +60
σ_0 .high	σ_0 .low	σ_0 .IP	σ_0 .SP	+20 / +40
Kdebug.high	Kdebug.low	Kdebug.entry	Kdebug.init	+10 / +20
KernDescPtr	APIFlags	API Version	$0_{(0/32)} \mid K_{(32)} \mid 4'_{(1)}$	+0
				+0
4C / +18	+8 / +10	+4 / +8	+0	

25

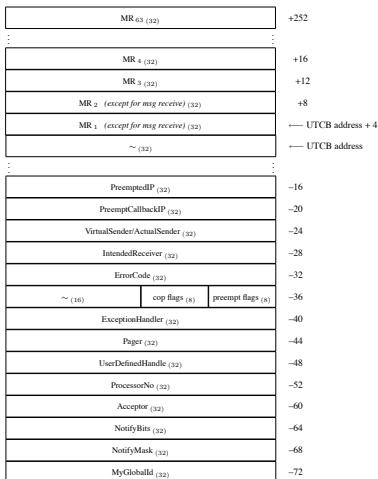
What's in the UTCB area?

- Every user thread has a User Thread Control Block (UTCB), which is a block of memory that the thread uses for communication with the kernel.
 - The UTCB contains:
 - Message registers (MRs)
 - Thread control registers (TCRs)
 - All UTCBs for a given address space are grouped in a single block called the UTCB area
 - Example: If UTCBs are 512 bytes long, then an address space with a 4KB UTCB area can support at most 8 threads

26

UTCB Layout (IA32)

- 64 Message “registers” named $MR_0, MR_1, \dots, MR_{63}$
 - Miscellaneous other fields:
 - ErrorCode
 - ExceptionHandler
 - Pager
 - Acceptor
 - ...
 - UTCB address points to the middle of the UTCB



2

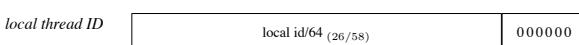
Trust, and UTCBs

- User processes can read and write whatever values they like in the UTCB (and in the UTCBs of other threads in the same address space)
 - Protected thread parameters (e.g., priority) must be stored in a separate TCB data structure that is only accessible to the kernel
 - Any data that is read from the UTCB cannot be trusted and must be validated by the kernel, as necessary, before use
 - Mappings for the UTCB area must be created by the kernel (otherwise user space code could cause the kernel to page fault by reading from an unmapped UTCB)

10

UTCB addresses and local thread ids

- Every UTCB must be 64-byte aligned, so the lower 6 bits in any UTCB address will be zero
 - Within a given address space, UTCB addresses are used as local thread ids:



- Other thread ids must have a nonzero value in their least significant 6 bits

29

How to find the UTCB

- Option 1: Design Protocol
 - User code assumes a predetermined UTCB address
 - Option 2: The UTCB pointer
 - At boot time, the kernel creates a 4 byte, read only segment in the GDT for a specific kernelspace address and loads a corresponding segment selector in %gs
 - The kernel stores the UTCB address of the current thread in that location
 - User code can read the UTCB address from %gs:0

30

Configuring an address space

- The addresses of the KIP and the UTCB can be set when a new address space is created:
- First, create a new thread in a new address space (we'll see how this is done soon)
- Now use the (privileged) SpaceControl system call:

<i>SpaceSpecifier</i>	EAX	<i>control</i>	EAX	<i>result</i>
<i>control</i>	ECX		ECX	~
<i>KernelInterfacePageArea</i>	EDX		EDX	~
<i>UtcbArea</i>	ESI		ESI	~
-	EDI		EDI	~
-	EBX		EBX	~
-	EBP		EBP	~
-	ESP		ESP	≡

call *SpaceControl*

- Threads cannot be activated (made runnable) until the associated address space has been configured in this way

31

Threads

32

Thread Ids

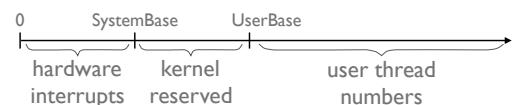
- User programs can reference other threads using thread ids

<i>global thread ID</i>	thread no (18/32)	version(14/32) ≠ 0 (mod 64)
<i>global interrupt ID</i>	intr no (18/32)	1 (14/32)
<i>local thread ID</i>	local id/64 (26/58)	000000
<i>nilthread</i>	0 (32/64)	
<i>anythread</i>	-1 (32/64)	
<i>anylocalthread</i>	-1 (26/58)	000000

33

Thread numbers

- Every thread number falls in to one of three ranges:



- The SystemBase and UserBase values are defined in the KIP
- Key insight: L4 translates hardware interrupts in to messages from (special) threads

34

Global ids bad ...

- The reliance on global ids is one of the weaknesses of the original L4 design
 - Any thread can reference any other thread by using its global id
 - Any thread can interfere with another thread (e.g., a denial of service attack) by using its global id
 - Even if thread ids are not officially published, they can still be guessed or faked
- We could avoid these problems if there were a way to ensure that any thread only had the **capability** to access a specific set of authorized threads ...

35

ThreadControl

- New threads are created using the (privileged) ThreadControl system call:
- Thread Control* →

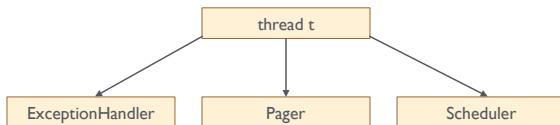
							≡
--	--	--	--	--	--	--	---

call *ThreadControl*
- If dest does not exist then the new thread is created in the same address space as SpaceSpecifier
 - If SpaceSpecifier=dest, then a new address space is created
 - The UTCBLocation must be within the UTCB area
 - If dest exists and SpaceSpecifier is nilthread, then the thread is deleted

36

Exception handlers, pagers, and schedulers

- Every thread has three associated threads



- The **exception handler** is responsible for dealing with any exceptions that t generates (specified in UTCB)
- The **pager** is responsible for dealing with any page faults that t generates (specified in UTCB)
- The **scheduler** is responsible for setting the priority and timeslice for t (hidden inside kernel TCB)

Schedule

- If s is the scheduler thread for t, then s can set t's scheduling parameters using the Schedule system call:

<i>dest</i>	EAX	→	EAX	<i>result</i>
—	ECX		ECX	~
—	EDX		EDX	~
processor control	EDX		ESI	~
preemption control	ESI		EDI	rem ts
ts len	EDI		EBX	rem total
total quantum	EBX		—	~
—	EBP		EBP	~
—	ESP		ESP	≡

- The specified priority cannot be higher than the scheduler's own priority
- ts is the timeslice: how long does the thread run before the kernel will switch to another thread
- quantum specifies a limit on the total time that a thread can run before it is suspended

37

38

ThreadSwitch

- A thread can give up any remaining part of its timeslice to another thread using the ThreadSwitch system call:

<i>dest</i>	EAX	→	EAX	≡
—	ECX		ECX	≡
—	EDX		EDX	≡
—	ESI		ESI	≡
—	EDI		EDI	≡
—	EBX		EBX	≡
—	EBP		EBP	≡
—	ESP		ESP	≡

call *ThreadSwitch*

- If dest is nilthread, then the caller still yields the CPU and the kernel determines which thread will run next ...

ExchangeRegisters

- A thread can read or write parameters of another thread using the ExchangeRegisters system call:

<i>dest</i>	EAX	→	EAX	<i>result</i>
<i>control</i>	ECX		ECX	<i>control</i>
<i>SP</i>	EDX		EDX	<i>SP</i>
<i>IP</i>	ESI		ESI	<i>IP</i>
<i>FLAGS</i>	EDI		EDI	<i>FLAGS</i>
<i>UserDefinedHandle</i>	EBX		EBX	<i>UserDefinedHandle</i>
<i>pager</i>	EBP		EBP	<i>pager</i>
—	ESP		ESP	≡

- ExchangeRegisters is not “privileged” ... but the destination thread must be in the same address space as the caller
- The exact effects of an ExchangeRegisters call are specified by a bit map in the control word:

<i>control</i>	from (18/32)	0 (3/19)	r d h p u f i s S R H
----------------	--------------	----------	-----------------------

39

40

IPC

IPC - Interprocess Communication

- IPC is a fundamental system call for communication between threads in L4
- A typical use of IPC proceeds as follows:
 - Load the message registers in the UTCB with a message to send
 - Invoke the IPC system call, which has two phases:
 - Send the message register values to a specified thread
 - Receive new message register values from a thread
 - Resume thread that initiated the IPC

41

42

Why combine send and receive phases?

- The combination of send and receive phases in a single system call:
 - requires only one system call instead of separate send and receive system calls
 - accomplishes both send and receive actions with only a single transition in to kernel mode
 - matches common communication idioms:
 - RPC: Send a request to a thread and wait for its reply
 - Server: Send response to a previous request and then wait for a new request to arrive

43

Synchronization and blocking

- Communication between threads requires a sender and a receiver
 - If either party is not ready, then the communication blocks
- Some versions of L4 allow an IPC call to specify timeout periods, after which a blocked IPC call will be aborted.
 - In practice, it is hard to come up with a good methodology for picking sensible timeout values
- Other versions of L4 support only two possible timeout options: 0 (non blocking) and ∞ (blocking)

44

Specifics

<i>to</i>	EAX	<i>from</i>
—	ECX	~
<i>FromSpecifier</i>	EDX	~
<i>MR</i> 0	ESI	<i>MR</i> 0
<i>UTCB</i>	EDI	≡
—	EBX	<i>MR</i> 1
—	EBP	<i>MR</i> 2
—	ESP	≡

— **Ipc** →

call Ipc

- Some message registers passed in CPU registers
- “to” can be nilthread, if there is no send phase
- “FromSpecified” can be:
 - nilthread, if there is no receive phase
 - anythread, if it is a server that will accept requests from any other thread

45

Message tags

- The value in *MR* 0 provides a message tag that describes the structure of the message in the remaining message registers:

<i>MsgTag</i> [<i>MR</i> 0]	label (16/48)	flags (4)	<i>t</i> (6)	<i>u</i> (6)

- label can be used to send/receive a 16 bit data value
- u* is the number of untyped words (uninterpreted 32 bit word values) sent in message registers
- t* is the number of typed-item words (MapItem, GrantItem; we'll talk about these soon ...)

46

Example: Interrupt handlers

- When a hardware interrupt occurs, the kernel sends an IPC message from the interrupt thread to its pager with the tag:

<i>From Interrupt Thread</i>	Initial SP (32/64)	Initial IP (32/64)	MR 2
	-1 (12/44)	0 (4)	MR 0

- When the pager has finished handling the error, it sends an IPC message back to the interrupt thread to reenable the corresponding interrupt

<i>To Interrupt Thread</i>	Initial SP (32/64)	Initial IP (32/64)	MR 1
	0 (16/48)	0 (4)	MR 0

47

Example: Thread start

- When a new thread is constructed, it waits for a message from its pager before starting:

<i>From Pager</i>	Initial SP (32/64)	Initial IP (32/64)	MR 2
	0 (16/48)	0 (4)	MR 0

- When a newly created thread receives a message of this form, the kernel loads the specified esp and eip values from the message in to the thread's context and marks the thread as being runnable ...

48

Example: Exception handling

- When a thread generates an exception, the kernel sends a message to the associated exception handler

EAX (32)	MR 12
ECX (32)	MR 11
EDX (32)	MR 10
ERX (32)	MR 9
ESP (32)	MR 8
EBP (32)	MR 7
ESI (32)	MR 6
EDI (32)	MR 5
ErrorCode (32)	MR 4
ExceptionNo (32)	MR 3
EFFLAGS (32)	MR 2
EIP (32)	MR 1
	MR 0
-4/-5 (12/14)	0 (6)
0 (6)	0 (6)
t = 0 (6)	u = 12 (6)

- If it chooses to resume the thread that generated the exception, it responds with a message of essentially the same format (possibly having updated registers in the process)

49

50

Address Space Management

Flexpages (fpages)

- A generalized form of “page” that can vary in size:

fpage (b, 2 ^b)	b/2 ¹⁰ (22/54)	s (6)	0 rwx
----------------------------	---------------------------	-------	-------

- Includes both 4KB pages and 4MB superpages as special cases
- Also includes special cases to represent the full address space (complete) and the empty address space (nilpage):

complete	0 (22/54)	s = 1 (6)	0 rwx
nilpage	0 (32/64)		

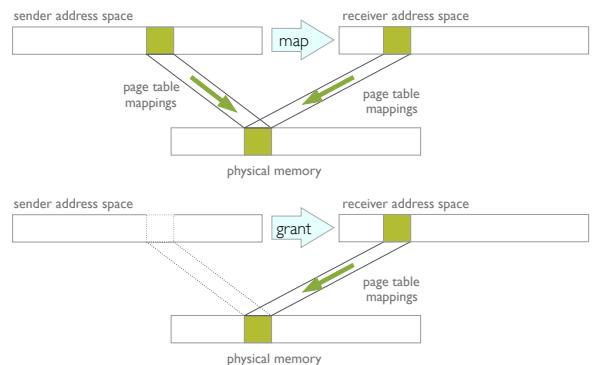
- Can be represented, in practice, using collections of 4KB and 4MB pages

51

52

Mapping and granting

- Address spaces in L4 are constructed by mapping or granting regions of memory between address spaces



MapItems and GrantItems

- A MapItem specifies a region of memory in the sender's address space that will be mapped into the receiver's address space

snd fpage (28/60)	0 rwx	MR $i+1$
snd base / 1024 (22/54)	0 (6)	10 00

- A GrantItem specifies a region of memory that will be removed from the sender's address space and added to the receiver's address space

snd fpage (28/60)	0 rwx	MR $i+1$
snd base / 1024 (22/54)	0 (6)	10 10

- Base values are used for mapping between fpages of different sizes; we will mostly ignore them for now

53

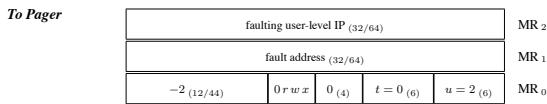
54

Typed items in IPC messages

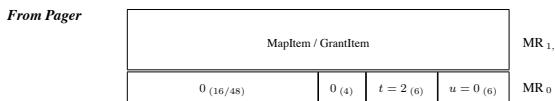
- An IPC message can contain multiple “typed items” (either MapItem or GrantItem values), that will create mappings in the receiver based on mappings in the sender
- The receiver sets an “acceptor” fpage in its UTCB to specify where newly received mappings should be received
- To receive anywhere, set the acceptor to “complete”
- To receive nowhere, set the acceptor to “nilpage”

Page faults

- When a thread triggers a page fault, the kernel translates that event into an IPC to the thread's pager:

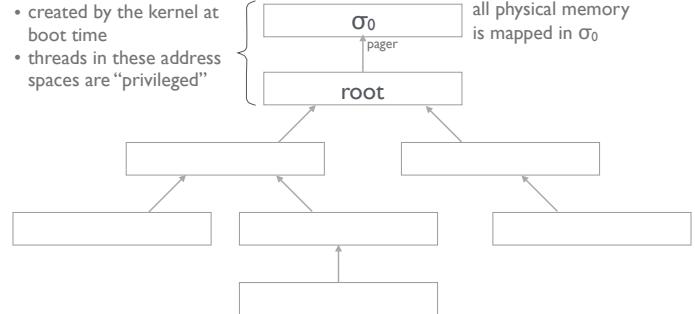


- The pager can respond by sending back a reply with a new mapping ... that also restarts the faulting thread:



The “recursive address space model”

- created by the kernel at boot time
- threads in these address spaces are “privileged”



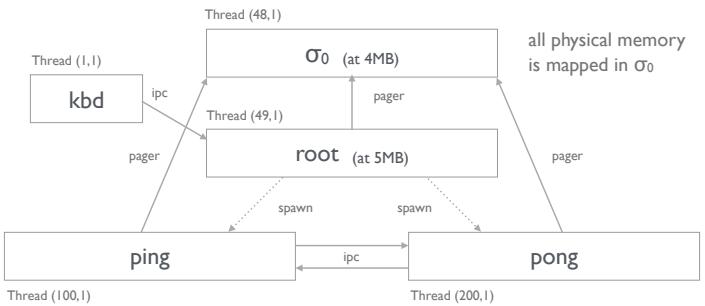
- In a dynamic system, we need the ability to revoke previous mappings ... this will get interesting ...

55

56

Let's look at an example ...

A demo using “pork”



57

58

Initialization code in root.c:

```
printf("This is a root server!\n");
showKIP();

ping = L4_GlobalId(100,1);
pong = L4_GlobalId(200,1);

//startPing();
spawn("ping", ping,      // Name & thread id
      0, ping,           // utcbNo & space spec
      L4_Myself(),       // Scheduler
      L4_Pager(),        // Pager
      (L4_Word_t)ping_thread, // eip
      ((L4_Word_t)pingstack) + PINGSTACKSIZE); // esp

//startPong();
spawn("pong", pong,      // Name & thread id
      0, pong,           // utcbNo & space spec
      L4_Myself(),       // Scheduler
      L4_Pager(),        // Pager
      (L4_Word_t)pong_thread, // eip
      ((L4_Word_t)pongstack) + PONGSTACKSIZE); // esp

//keyboard listener
L4_Threadid_t keyId = L4_GlobalId(1, 1); // Keyboard on IRQ1
L4_Threadid_t rootId = L4_MyglobalId(); // My id

printf("keyboard id = %x, my id = %x\n", keyId, rootId);
printf("associate produces %x\n",
L4_AssociateInterrupt(keyId, rootId));
```

Event loop code in root.c:

```
L4_MsgTag_t tag = L4_Receive(keyId);
for (;;) {
    printf("received msg (tag=%x) from %x\n", tag, keyId);
    if (L4_IpcSucceeded(tag) &&
        L4_UntypedWords(tag)==0 &&
        L4_TypedWords(tag) ==0) {
        L4_LoadMR(0, 0); // tag: Empty message, ping back to interrupt thread
        tag = L4_Call(keyId);
        printf("root's Call completed ...\n");
    } else {
        printf("Ignoring message/failure, trying again ...\n");
        tag = L4_Receive(keyId);
    }
}
printf("This message won't appear!\n");
```

59

60