COMP9242 Advanced OS

S2/2017 W06: **Real-Time Systems**

@GernotHeiser

Incorporating material by Stefan Petters

Never Stand Still | Engineering | Computer Science and Engineering

Copyright Notice

These slides are distributed under the Creative Commons Attribution 3.0 License

- You are free:
  - to share—to copy, distribute and transmit the work
  - to remix—to adapt the work
- under the following conditions:
  - **Attribution:** You must attribute the work (but not in any way that suggests that the author endorses you or your use of the work) as follows:
    "Courtesy of Gernot Heiser, UNSW Sydney"

The complete license text can be found at http://creativecommons.org/licenses/by/3.0/legalcode

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Real-Time Basics

 © 2017 Gernot Heiser. Distributed under CC Attribution License

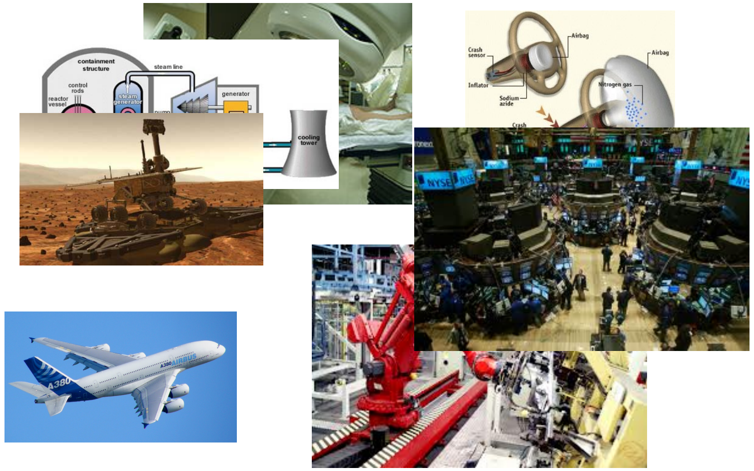
Real-Time System: Definition

A real-time system is any information processing system which has to respond to externally generated input stimuli within a finite and specified period

- Correctness depends not only on the logical result (function) but also the time it was delivered
- Failure to respond is as bad as delivering the wrong result!

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Real-Time Systems



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Types of Real-Time Systems

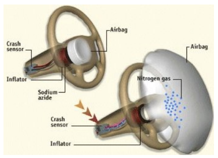
- Hard real-time systems
- Weakly-hard real-time systems
- Firm real-time systems
- Soft real-time systems
- Best-effort systems

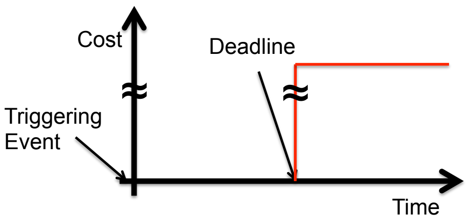
- Real-time systems typically deal with *deadlines*:
  - A deadline is a time instant by which a response has to be completed
  - A deadline is usually specified as *relative* to an event
    - The *relative deadline* is the *maximum allowable response time*
    - Absolute deadline: event time + relative deadline

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Hard Real-Time Systems

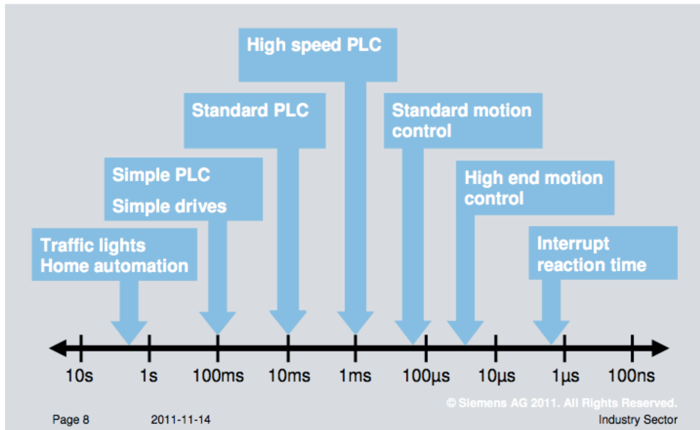
- Deadline miss is "catastrophic"
  - safety-critical system: failure results in death, severe injury
  - mission-critical system: failure results in massive financial damage
- Steep and real "cost" function





 © 2016 Gernot Heiser. Distributed under CC Attribution License

Eg RT Requirements in Industrial Automation



Source: Siemens

 © 2016 Gernot Heiser. Distributed under CC Attribution License

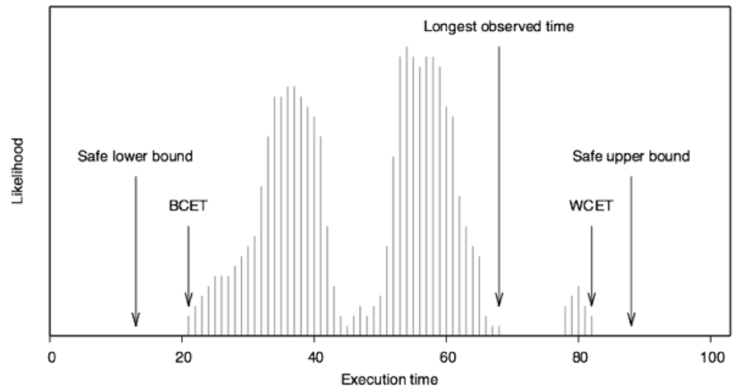
Real-Time ≠ Real Fast

System	Deadline	Single Miss Conseq	Ultimate Conseq.
Car engine ignition	2.5 ms	Catastrophic	Engine damage
Industrial robot	5 ms	Recoverable?	Machinery damage
Air bag	20 ms	Catastrophic	Injury or death
Aircraft control	50 ms	Recoverable	Crash
Industrial process	100 ms	Recoverable	Lost production, plant/ environment damage
Pacemaker	100 ms	Recoverable	Death

Challenge of real-time systems: **Guaranteeing** deadlines

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Challenge: Execution-Time Variance



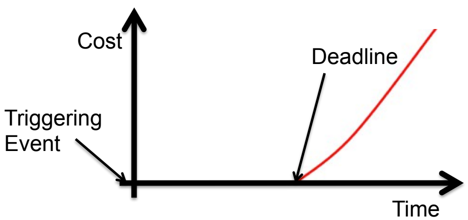
Variance may be orders of magnitude!

- Data-dependent execution path
- Micro-architectural features: pipelines, caches

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Weakly-Hard Real-Time Systems

- Tolerate a (small) fraction of deadline misses
  - Most feedback control systems (including life-supporting ones!)
    - occasionally missed deadline can be compensated at next event
    - system becomes unstable if too many deadlines are missed
  - Typically integrated with other fault tolerance
    - electro-magnetic interference, other hardware issues

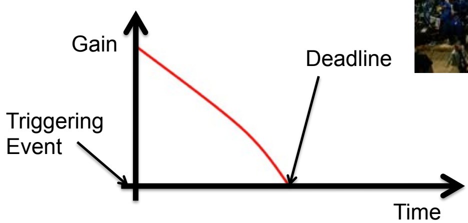


In practice, certifiers treat avionics as hard RT

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Firm Real-Time Systems

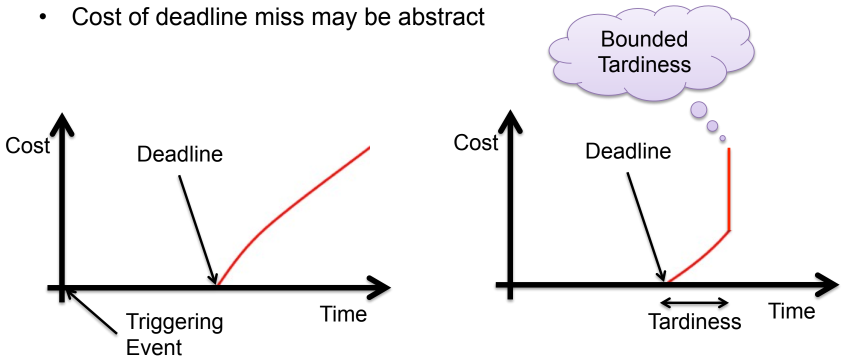
- Deadline miss makes computation obsolete
  - Typical examples are forecast systems
    - weather forecast
    - trading systems
- Cost may be loss of revenue (gain)



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Soft Real-Time Systems

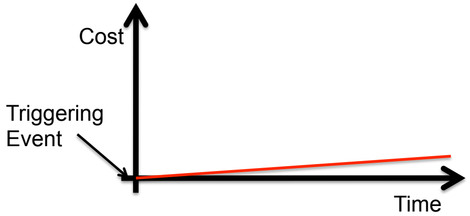
- Deadline miss is undesired but tolerable
  - Frequently results on quality-of-service (QoS) degradation
    - eg audio, video rendering
    - Steep "cost" function
- Cost of deadline miss may be abstract



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Best-Effort Systems

- No deadlines, timeliness is not part of required operation
- In reality, there is at least a nuisance factor to excessive duration
  - response time to user input
- Again, "cost" may be reduced gain



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Real-Time Operating System (RTOS)

- Designed to support real-time operation
  - Fast context switches, fast interrupt handling?
  - Yes, but *predictable* response time is more important
    - "Real time is not real fast"
  - Analysis of *worst-case execution time* (WCET)
- Support for *scheduling policies* appropriate for real time
- Classical RTOSes very primitive
  - single-mode execution
  - no memory protection
  - essentially a scheduler with a threads package
  - "real-time executive"
  - inherently cooperative
  - inherently *trust all code*
- Many modern uses require actual OS technology for isolation
  - generally microkernels
  - QNX, Integrity, VXworks, L4 kernels

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Approaches to Real Time

- Clock-driven (cyclic)
  - *Periodic scheduling*
  - Typical for control loops
  - Fixed order of actions, round-robin execution
  - *Statically* determined (static schedule) if periods are fixed
    - need to know all execution parameters at system configuration time

> **Emulation on event-driven system:** treat clock tick as event

> **Emulation on clock-driven system**: buffer event (IRQ) until timer tick

- Event-driven
  - *Sporadic scheduling*
  - Typical for reactive systems (sensors & actuators)
  - Static or dynamic schedules
  - Analysis requires bounds on event arrivals

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Real-Time System Operation

- Time-triggered
  - Pre-defined temporal relation of events
  - event is not serviced until its defined *release time* has arrived
- Event-triggered
  - timer interrupt
  - asynchronous events
- Rate-based
  - activities get assigned CPU shares ("*rates*")

© 2016 Gernot Heiser. Distributed under CC Attribution License

Real-Time Task Model

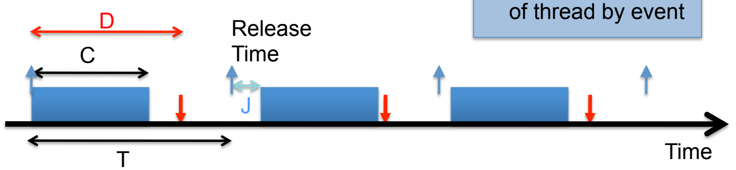
- **Job**: unit of work to be executed
  - … resulting from an event or time trigger
- **Task**: set of related jobs which provide some system function
  - A *task* is a sequence of *jobs* (typically executing same function)
  - Job *i+1* of of a task cannot start until job *i* is completed/aborted
- Periodic tasks
  - Time-driven and all relevant characteristics known a priori
    - Task t characterized by period T_i, deadline, D_i and execution time C_i
    - Applies to all jobs of task
- Aperiodic tasks
  - Event driven, characteristics are not known a priori
    - Task t characterized by period T_i, deadline D_i and arrival distribution
- Sporadic tasks
  - Aperiodic but with known minimum inter-arrival time T_i
  - treated similarly to periodic task with period T_i

© 2016 Gernot Heiser. Distributed under CC Attribution License

Standard Task Model

C: Worst-case computation time (WCET)
T: Period (periodic) or minimum inter-arrival time (sporadic)
D: Deadline (relative, frequently "implicit deadlines" D=T)
J: Release jitter
P: Priority: higher number means higher priority
B: Worst-case blocking time
R: Worst-case response time
U: Utilisation; U=C/T

OS terminology:
- "task" = thread
- "job" = unblocking of thread by event



© 2016 Gernot Heiser. Distributed under CC Attribution License

Task Constraints

- Deadline constraint: must complete before deadline
- Resource constraints:
  - Shared (R/O), exclusive (W-X) access
  - Energy
  - Precedence constraints:
    $t_1 \Rightarrow t_2$: t_2 execution cannot start until t_1 is finished
  - Fault-tolerance requirements
    - eg redundancy
- Scheduler's job to ensure that constraints are met!

Locking!

© 2016 Gernot Heiser. Distributed under CC Attribution License

Scheduling

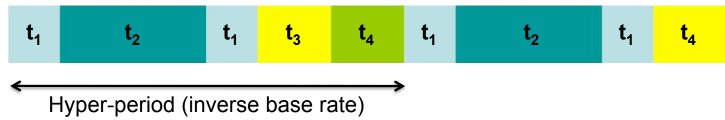
- Preemptive vs non-preemptive
- Static (fixed, off-line) vs dynamic (on-line)
- Clock-driven vs priority-based
  - clock-driven is static, only works for very simple systems
  - priorities can be static (pre-computed and fixed) or dynamic
  - dynamic priority adjustment can be at task-level (each job has fixed prio) or job-level (jobs change prios)

© 2016 Gernot Heiser. Distributed under CC Attribution License

Clock-Driven (Time-Triggered) Scheduling

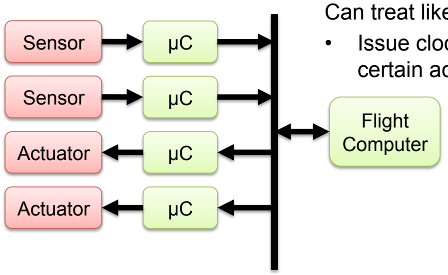
- Typically implemented as time "frames" adding up to "base rate"
- Advantages
  - fully deterministic
  - "cyclic executive" is trivial
  - minimal overhead
- Disadvantage:
  - Big latencies if event rate doesn't match base rate (hyper-period)
  - Inflexible

```
while (true) {
   wait_tick();
   job_1();
   wait_tick();
   job_2();
   wait_tick();
   job_1();
   wait_tick();
   job_3();
   wait_tick();
   job_4();
}
```



© 2016 Gernot Heiser. Distributed under CC Attribution License

Synchronous Distributed RT Systems



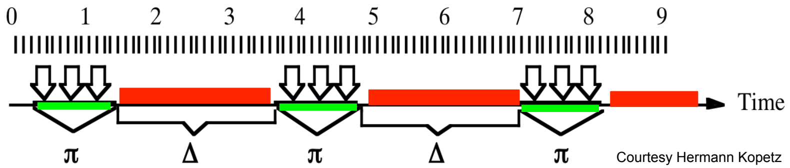
Can treat like single system if clocks synchronised?

- Issue clock drift: can only synchronise within certain accuracy

Time-triggered architecture

Idea: use *sparse time*:

- Restrict events to *active interval* π
- Separated by *silence interval* Δ
- Δ allows for clock drift and communications time



Courtesy Hermann Kopetz

© 2016 Gernot Heiser. Distributed under CC Attribution License

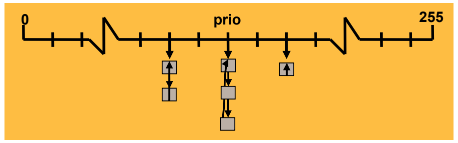
Non-Preemptive Scheduling

- Minimises context-switching overhead
  - Significant cost on modern processors (pipelinies, caches)
- Easy to analyse timeliness
- Drawbacks:
  - Larger response times for "important" tasks
  - Reduced utilisation, schedulability
    - In many cases cannot produce schedule despite plenty idle time
  - Can't re-use slack (eg for best-effort)
- Only used in very simple systems

© 2016 Gernot Heiser. Distributed under CC Attribution License

Fixed-Priority Scheduling (FPS)

- Real-time priorities are absolute:
  - Scheduler always picks highest-priority job
- Obviously easy to implement, low overhead
- Drawbacks: inflexible, sub-optimal
  - Cannot schedule some systems which are schedulable preemptively
- Note: “Fixed” prios in the sense that system doesn’t change them
  - OS may support dynamic adjustment
  - Requires on-the-fly (re-)admission control



Classical L4 scheduling
- Hard thread priority
- Round-robin within prio
- Time slice for preemption

Choosing Prios: Rate-Monotonic Scheduling (RMS)

- RMS: Standard approach to *fixed priority assignment*
  - $T_i<T_j \Rightarrow P_i>P_j$
  - 1/T is the “rate” of a task
- RMS is *optimal* for fixed priorities
- Schedulability test: RMS can schedule n tasks with D=T if

  $U \equiv \sum C_i/T_i \leq n(2^{1/n}-1)$; $\lim_{n\to\infty} U = \log 2$

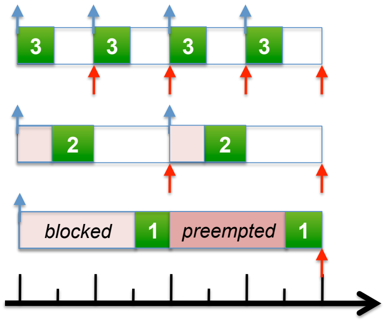
n	1	2	3	4	5	10	∞
U [%]	100	82.8	78.0	75.7	74.3	71.8	69.3

- If D<T replace by *deadline-monotonic scheduling* (DMS):
  - $D_i<D_j \Rightarrow P_i>P_j$
- DMS is also optimal (but schedulability bound is more complex)

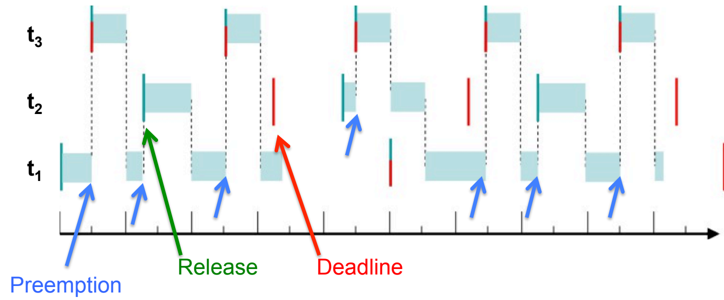
Rate-Monotonic Scheduling

RMS schedulability condition is sufficient but not necessary

	T	D	P	C	U [%]
t_3	20	20	3	10	50
t_2	40	40	2	10	25
t_1	80	80	1	20	25
					100



FPS Example



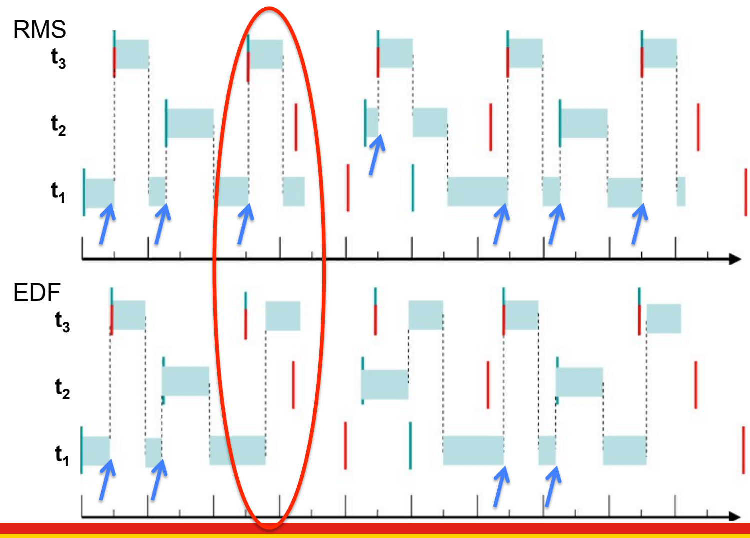
	P	C	T	D	U [%]	release
t_3	3	5	20	20	25	5
t_2	2	8	30	20	27	12
t_1	1	15	50	50	30	0
					82	

Choosing Prios: Earliest Deadline First (EDF)

- *Dynamic scheduling policy*
- Job with closest deadline executes
- Preemptive EDF with D=T is *optimal*: n jobs can be scheduled iff
  $U \equiv \sum C_i/T_i \leq 1$
  - necessary and sufficient condition
  - no easy test if D≠T

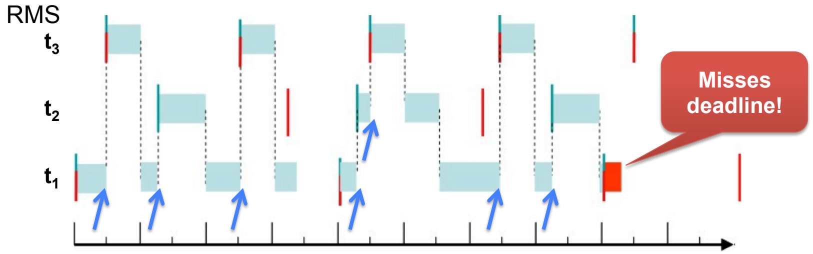
 © 2016 Gernot Heiser. Distributed under CC Attribution License

FPS vs EDF



 © 2016 Gernot Heiser. Distributed under CC Attribution License

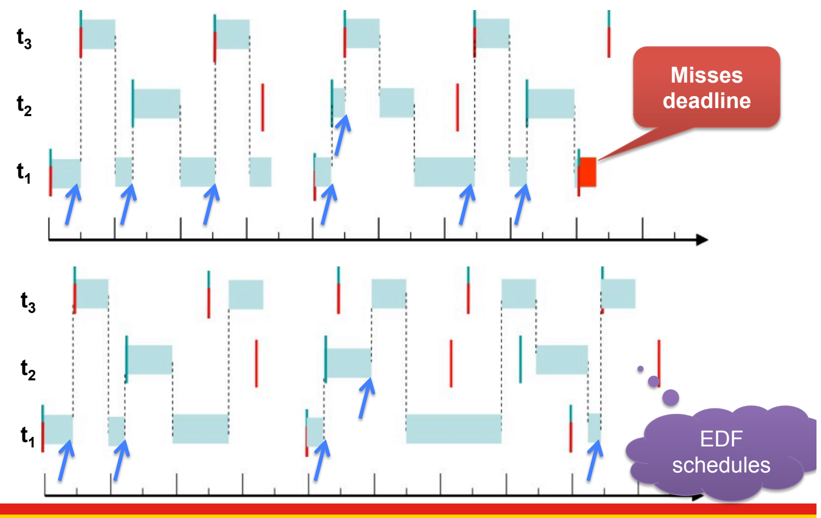
FPS vs EDF



	P	C	T	D	U [%]	release
t_3	3	5	20	20	25	5
t_2	2	8	30	20	27	12
t_1	1	15	40	40	37.5	0
					89.5	

 © 2016 Gernot Heiser. Distributed under CC Attribution License

FPS vs EDF

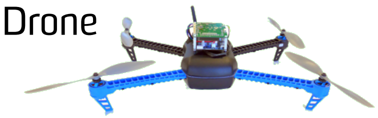


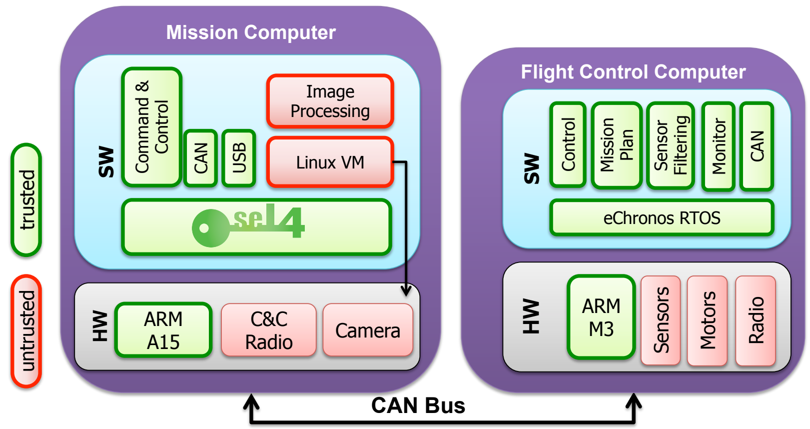
 © 2016 Gernot Heiser. Distributed under CC Attribution License

Resource Sharing

 © 2017 Gernot Heiser. Distributed under CC Attribution License

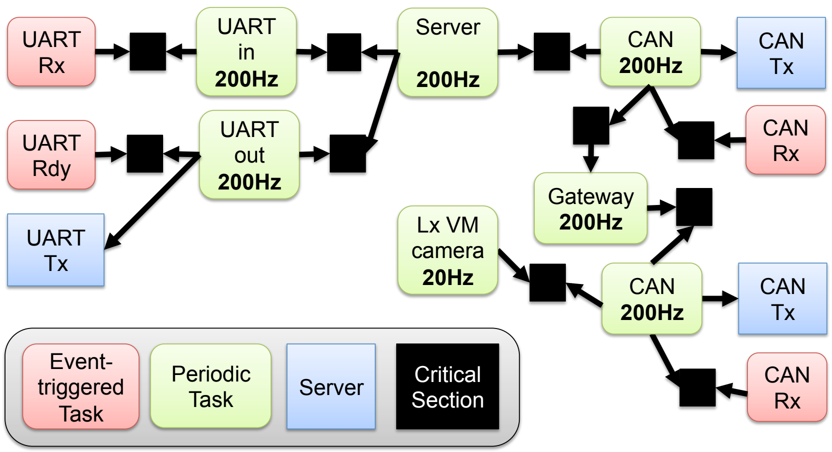
Example: SMACCMcopter Drone





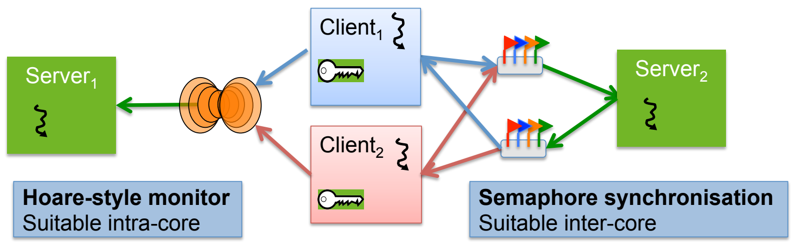
 © 2016 Gernot Heiser. Distributed under CC Attribution License

SMACCMcopter Mission Computer Architecture



 © 2016 Gernot Heiser. Distributed under CC Attribution License

seL4 Sharing: Critical Sections as Servers



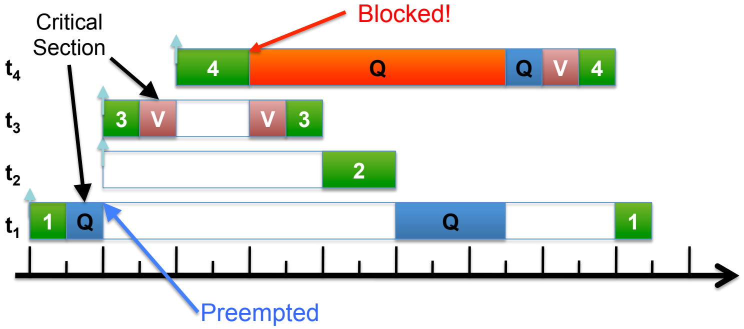

```
serv_local() {
  …
  wait(ep);
  while (1) {
    /* critical section */
    Reply&wait(ep);
  }
}
```

```
client() {
  while (1) {
    …
    call(ep);
    …
    signal(eap_ry);
    …
    wait(eap_rq);
  }
}
```

```
Serv_remote() {
  …
  while (1) {
    wait(eap_rq);
    /* critical section */
    signal(eap_ry);
  }
}
```

 © 2016 Gernot Heiser. Distributed under CC Attribution License

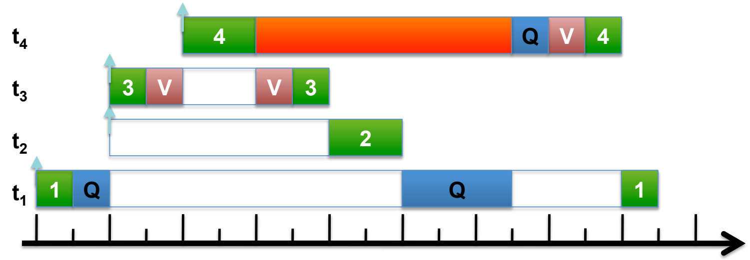
Problem: Priority Inversion

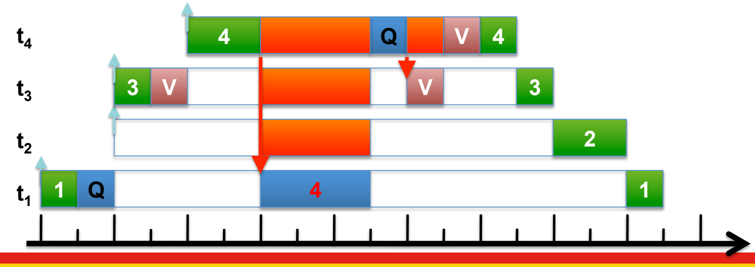


- High-priority job is blocked for a long time by a low-prio job
- Long wait chain: $t_1 \rightarrow t_4 \rightarrow t_3 \rightarrow t_2$
- Worst-case blocking time of t_1 bounded by total WCET: $C_2+C_3+C_4$
- Must find a way to do better!

 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Inheritance ("Helping")

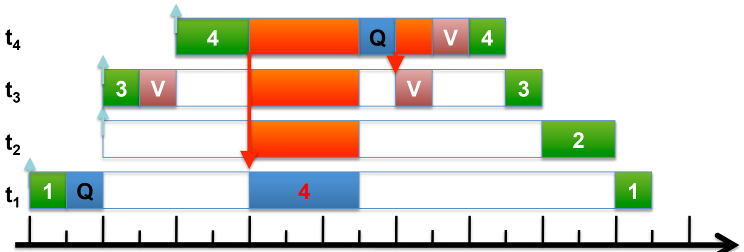




 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Inheritance

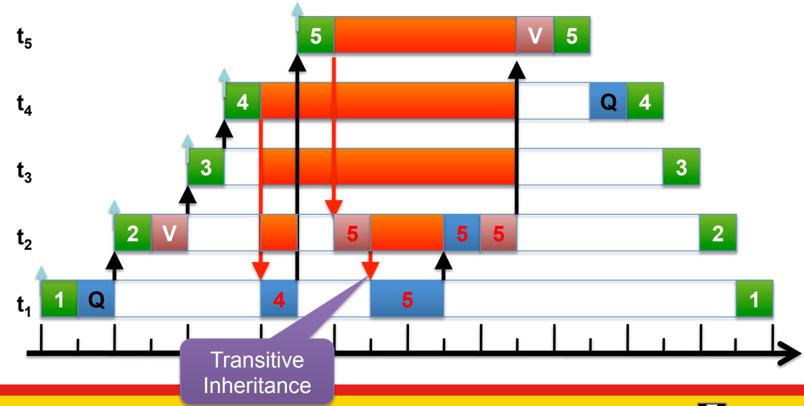
- If t_1 blocks on a resource held by t_2, *and* $P_1>P_2$, then
  - t_2 is temporarily given priority P_1
  - when t_t releases the resource, its priority reverts to P_2



 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Inheritance

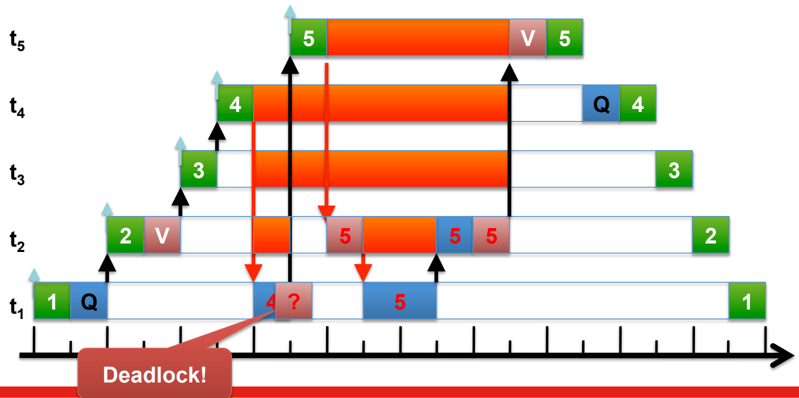
- If t_1 blocks on a resource held by t_2, *and* $P_1>P_2$, then
  - t_2 is temporarily given priority P_1
  - when t_t releases the resource, its priority reverts to P_2



 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Inheritance

- If t_1 blocks on a resource held by t_2, *and* $P_1 > P_2$, then
  - t_2 is temporarily given priority P_1
  - when t_t releases the resource, its priority reverts to P_2



 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Inheritance Protocol (PIP)

- If t_1 blocks on a resource held by t_2, *and* $P_1 > P_2$, then
  - t_2 is temporarily given priority P_1
  - when t_t releases the resource, its priority reverts to P_2
- Transitive inheritance
  - potentially long blocking chains
  - potential for deadlock
- Frequently blocks much longer than necessary

Priority Inheritance:
- Easy to use
- Potential deadlocks
- Complex to implement
- Bad worst-case blocking times

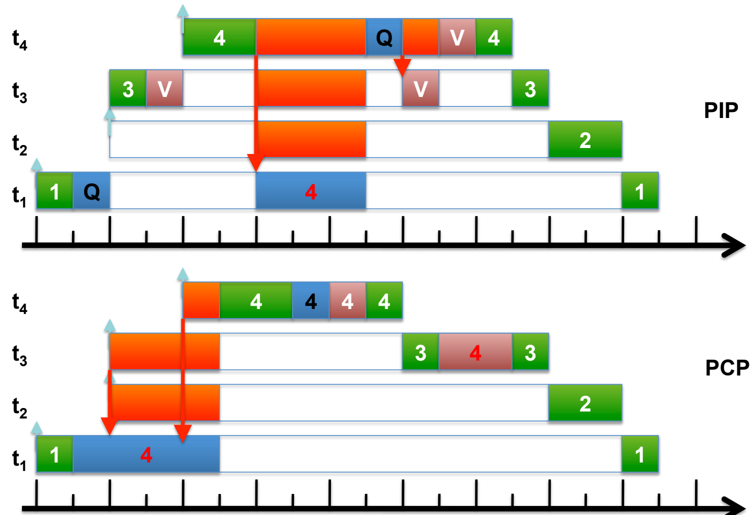
 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

Priority Ceiling Protocol (PCP)

- Purpose: ensure job can block at most once on a resource
  - avoid transitivity, potential for deadlocks
- Idea: associate a *ceiling priority* with each resource
  - equal to the highest priority of jobs that may use the resource
  - when job accesses its resource, immediately bump prio to ceiling!
- Also called:
  - *immediate ceiling priority protocol* (ICPP)
  - *ceiling priority protocol* (CPP)
  - *stack-based priority-ceiling protocol*
    - because it allows running all jobs on the same stack (i.e. thread)
- Improved version of the *original ceiling priority protocol* (OCPP)
  - … which is also called the *basic priority ceiling protocol*
  - OPCP bumps prio to ceiling of all *waiting* threads
  - IPCP requires global tracking of ceiling prios

 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

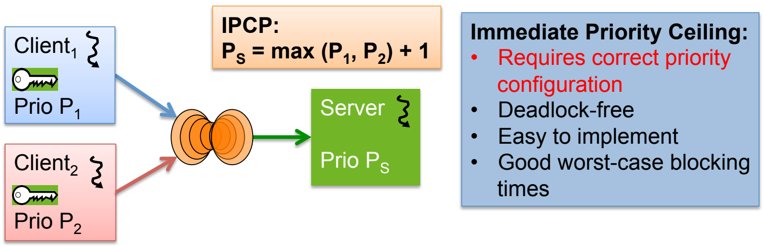
(Immediate) Priority Ceiling Protocol



 © 2016 Gernot Heiser. Distributed under CC Attribution License UNSW AUSTRALIA

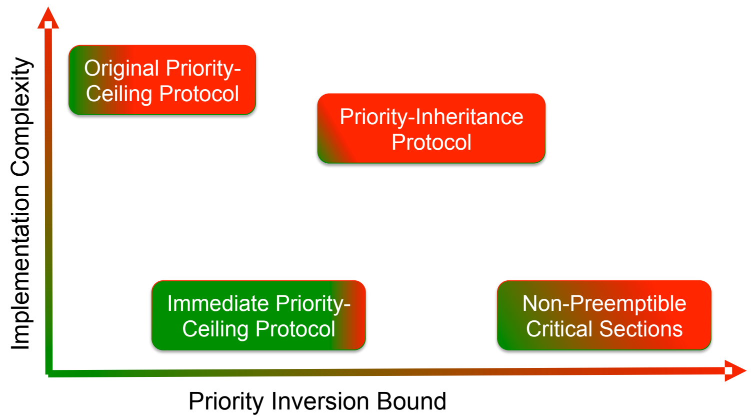
IPCP Implementation

- Each task must declare all resources at admission time
  - System must maintain list of tasks associated with resource
  - Priority ceiling derived from this list
  - For EDF the "ceiling" is the *floor of relative deadlines*
- seL4: "resource declaration" is implicit in capability distribution
  - Using critical section requires cap for server's request endpoint



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Comparison of Locking Protocols



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Scheduling Overloaded RT Systems

 © 2017 Gernot Heiser. Distributed under CC Attribution License

Naïve Assumptions: Everything Schedulable

Standard assumptions of classical RT systems:

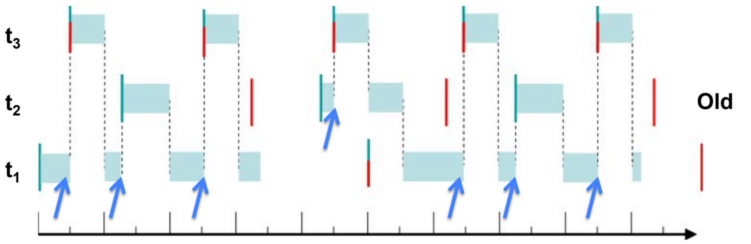
- All WCETs known
- All jobs complete within WCET
- Everything is Trusted

What happens if those assumptions are not met?

- **Overloaded system**:
  - Not all tasks complete within WCET or
  - Total utilisation exceeds schedulability bounds
- System no longer schedulable – what loses?

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Overload: FPS

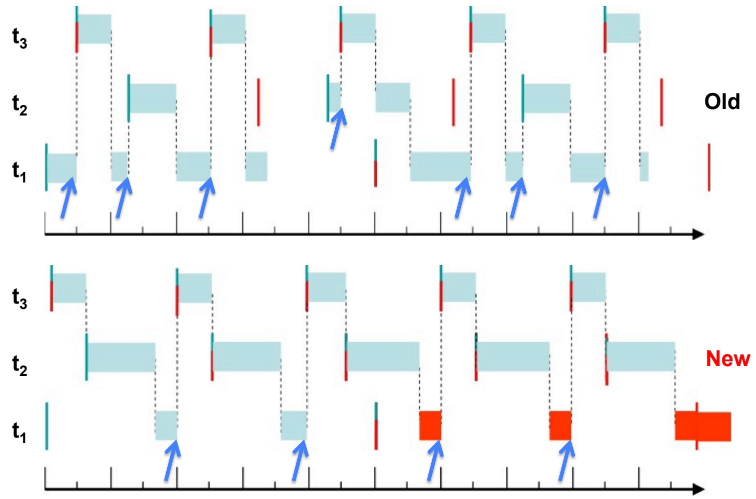


New

	P	C	T	D	U [%]
t_3	3	5	20	20	25
t_2	2	12	20	20	60
t_1	1	15	50	50	30
					115

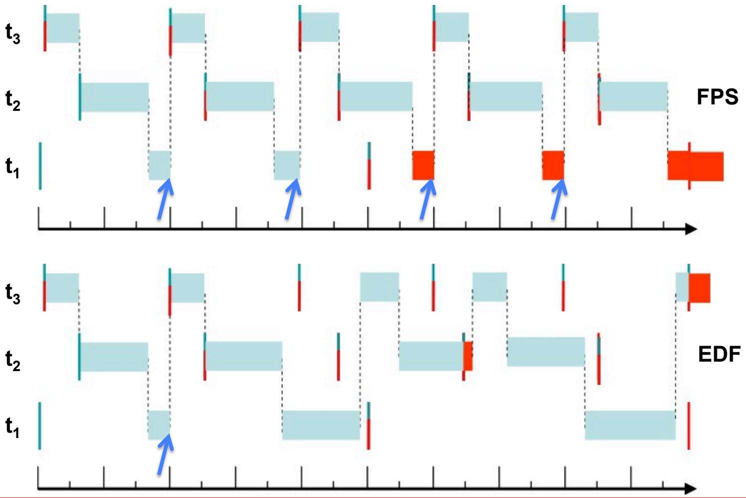
 © 2016 Gernot Heiser. Distributed under CC Attribution License

Overload: FPS



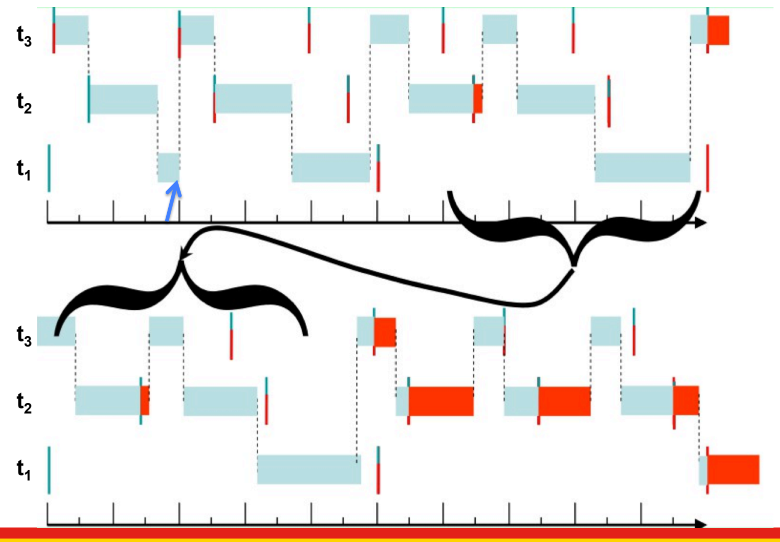
 © 2016 Gernot Heiser. Distributed under CC Attribution License

Overload: FPS vs EDF



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Overload: EDF



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Overload: FPS vs EDF

- **On overload, (by definition!)** ***lowest-prio jobs miss deadlines***
- Result is well-defined and -understood for FPS
  - Treats highest-prio task as “most important”
  - … but that may not always be appropriate!
  - Under transient overload may miss deadlines of higher-priority tasks
- Result is unpredictable (seemingly random) for EDF
  - May result in all tasks missing deadlines!
  - Under constant overload will scale back all tasks
  - No concept of task “importance”
  - “EDF behaves badly under overload”
  - Main reason EDF is unpopular in industry

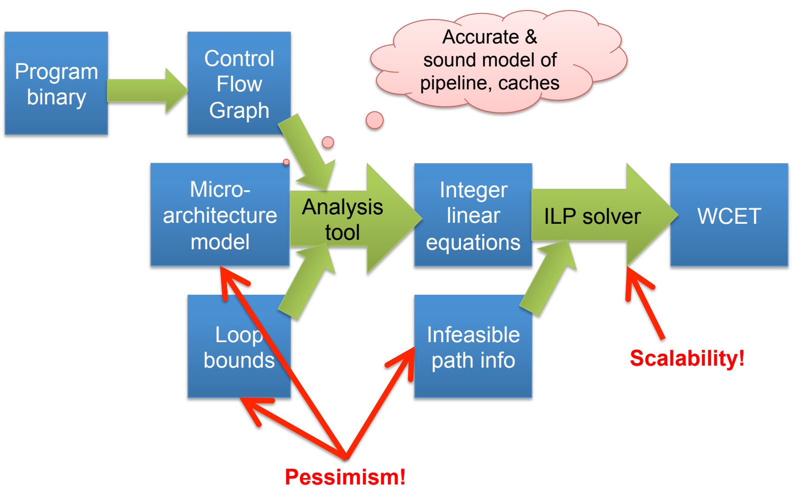
 © 2016 Gernot Heiser. Distributed under CC Attribution License

Why Have Overload?

- Faults (software, EMI, hardware)
- Incorrect assumptions about environment
- Optimistic WCET
  - Computing WCET of non-trivial programs is hard, often infeasible!
  - Safe WCET bounds tend to be highly pessimistic (orders of magnitude!)
  - WCET often very unlikely and orders of magnitude worse than “normal”
    - Estimation inaccuracies from caches, pipelines, under-specified hardware…
    - “normal” vs “exceptional” operating conditions
    - requires massive over-provisioning
  - Some systems have effectively unbounded execution time
    - e.g. object tracking

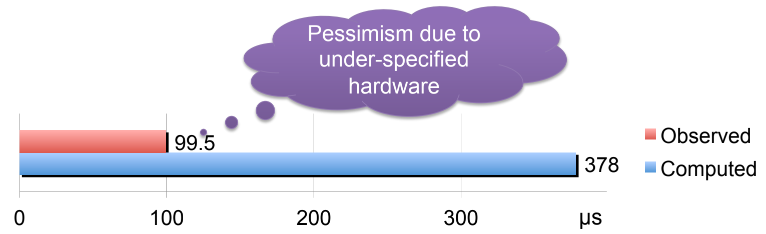
 © 2016 Gernot Heiser. Distributed under CC Attribution License

WCET Analysis



 © 2016 Gernot Heiser. Distributed under CC Attribution License

seL4 WCET Analysis on ARM11



WCET presently limited by verification practicalities
- **without regard to verification achieved 50 µs**
- **10 µs seem achievable**
- **BCET ~ 1µs**
- **[Blackham‘11, ‘12] [Sewell’16]**

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Why Overload? SWaP Challenge

Traditional embedded-systems approach: one µ-controller per function

- Automotive reached 100 ECUs in top-of-line cars 10 years ago
- ECUs must be robust – expensive
  - Tolerant to wide temperature range
  - Resistant to dust, water, grease, acid
  - Resistant to Vibrations
- **SWaP: space, weight and power**
  - Overhead of packaging, cabling
- Autonomous vehicles require even more functions than traditional
  - Also integration/cooperation of functionality
  - Infotainment/driver assist: consumer electronics + automotive control
- General challenge for cyber-physical systems (CPS)
  - Robots, autonomous aircraft, smart factories

Forces consolidation of multiple functions on single processor

© 2017 Gernot Heiser. Distributed under CC Attribution License

Why Have Overload?

- Faults (software, EMI, hardware)
- Incorrect assumptions about environment
- Optimistic WCET
  - Computing WCET of non-trivial programs is hard, often infeasible!
  - Safe WCET bounds tend to be highly pessimistic (orders of magnitude!)
  - WCET often very unlikely and orders of magnitude worse than "normal"
    - thanks to caches, pipelines, under-specified hardware
    - requires massive over-provisioning
- *Consolidation of functionality*

Way out?

- Need explicit notion of importance: *criticality*
- Expresses effect of failure on the system mission
  - Catastrophic, hazardous, major, minor, no effect
- *Orthogonal to scheduling priority!*

© 2016 Gernot Heiser. Distributed under CC Attribution License

Mixed-Criticality Systems

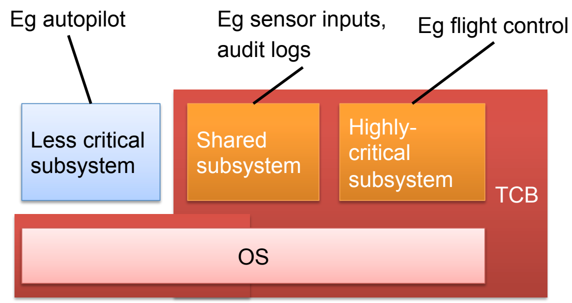
© 2017 Gernot Heiser. Distributed under CC Attribution License

Consolidation: Mixed-Criticality Systems (MCS)

Certification requirement: More critical components must *not* depend on any less critical ones! [ARINC-653]

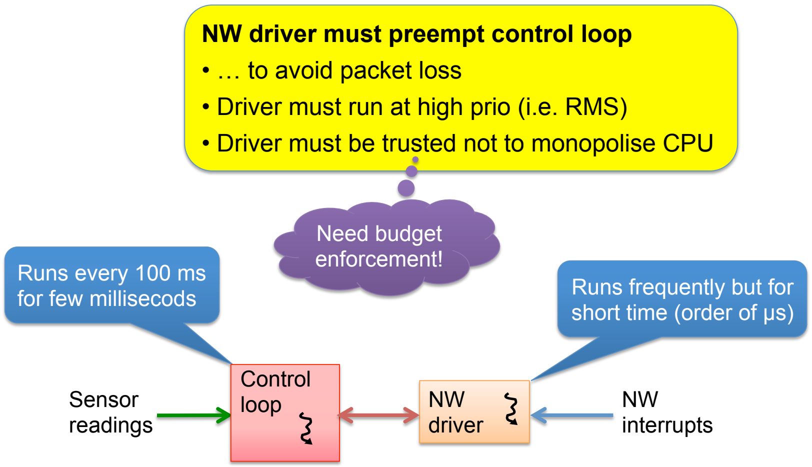
Higher criticality certification
- **More expensive**
- **More pessimistic WCET**

⇒ Minimise high-crit!



© 2016 Gernot Heiser. Distributed under CC Attribution License

Criticality = Importance ≠ Priority



 © 2016 Gernot Heiser. Distributed under CC Attribution License

OS Support For Mixed Criticality

MCS need strong OS-enforced isolation

- Spatial isolation: memory protection
  - Address space
- Temporal isolation: enforce CPU time limit enforcement
  - Time budget
- Criticality notion:
  - Stop **LOW** from overrunning budget
  - Get out of jail if **HIGH** overruns optimistic budget
    - Must be fast, as the cost of change must be included in analysis!

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Mixed Criticality Example

Criticality	T	U_{HIGH}	U_{MED}	U_{LOW}	$U_{average}$
High	10	50%	20%	20%	0.05%
Medium	1	N/A	60%	20%	2.5%
Low	100	N/A	N/A	unknown	10%
Total		**50%**	**80%**	**over**	**12.55%**

- **HIGH** alone has poor utilisation ⇒ gain from consolidation
- **HIGH+MEDIUM** can be scheduled for med-crit WCET
- **HIGH+MEDIUM cannot** be scheduled for most conservative WCET
- Idea: schedule under optimistic assumptions
  - Prioritise **HIGH** if it overruns its **MEDIUM** WCET
  - Change **HIGH** budget to pessimistic value

 © 2016 Gernot Heiser. Distributed under CC Attribution License

Budgets in EDF: CPU Bandwidth Reservations

- Idea: Utilisation U = C/T can be seen as required CPU *bandwidth*
  - Account time use against reservation C
  - Not runnable when reservation exhausted
  - Replenish every T
- Can support over-committing
  - Reduce **LOW** reservations if **HIGH** reservations fully used
- Advantages:
  - Allows dealing with jobs with unknown (or untrusted) deadlines
  - Allows integrating sporadic, asynchronous and soft tasks
- Modelled as a "server" which hands out time to jobs
  - effectively a simple (FIFO) sub-scheduler
  - *Constant-bandwidth server* (CBS) [Abeni & Buttazzo '98]

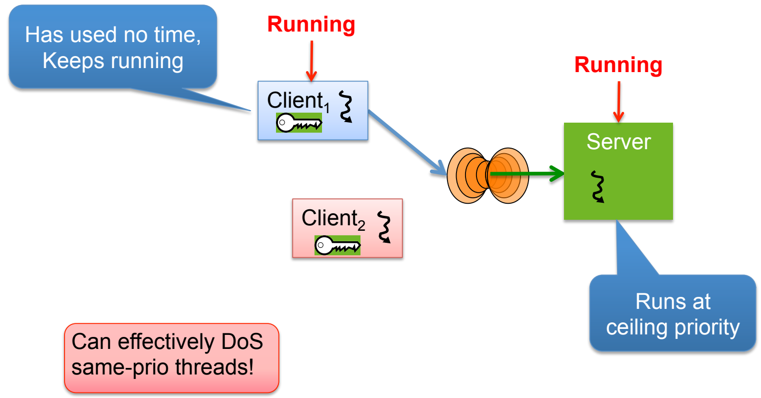
 © 2016 Gernot Heiser. Distributed under CC Attribution License

Mixed Criticality Implementation

- Whenever running **LOW** job, ensure no **HIGH** job misses deadline
- Switch to *critical mode* when not assured
  - Various approaches to determine switch
  - eg. *zero slack*: **HIGH** job's deadline = its WCET
- Criticality-mode actions:
  - FP: temporarily raise all **HIGH** jobs' prios above that of all others
    - Simply preempting present job won't help!
  - EDF: drop all **LOW** deadlines earlier than next **HIGH** deadline
- Issues:
  - Treatment of **LOW** jobs still rather indiscriminate:
    - EDF: switch will force deadline misses
    - FPS: **LOW** gets second chance
  - Need to determine when to switch to normal mode, restore prios
  - Switch must be fast – must be allowed for in schedulability analysis!

 © 2016 Gernot Heiser. Distributed under CC Attribution License

seL4 Challenge: Shared Servers



 © 2016 Gernot Heiser. Distributed under CC Attribution License

seL4 Separate Scheduling Properties & Threads

Classical Thread Attributes

- Priority
- Time slice

Not runnable if null

New Thread Attributes

- Priority
- Scheduling context capability

Upper bound, not reservation!

Scheduling context object
- T: period
- C: budget (≤ T)

Not yet in mainline!

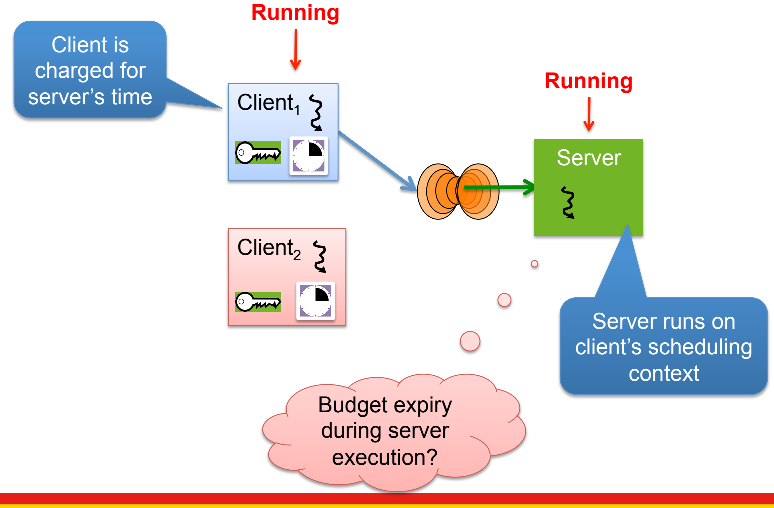
C = 2
T = 3

C = 250
T = 1000

`SchedControl` capability conveys right to assign budgets (i.e. perform admission control)

 © 2016 Gernot Heiser. Distributed under CC Attribution License

seL4 Shared Server with Scheduling Contexts



 © 2016 Gernot Heiser. Distributed under CC Attribution License

Budget Expiry Options

- Multi-threaded servers (COMPOSITE [Parmer '10])
  - Forcing all servers to be thread-safe is policy 😢
  - Optional in seL4 model
- Bandwidth inheritance with "helping" (Fiasco [Steinberg '10])
  - Ugly dependency chains 😢
  - Wrong thread charged for recovery cost 😢
- Use *timeout exceptions* to trigger one of several possible actions:
  - Provide emergency budget to leave critical section
  - Cancel operation & roll-back server
  - Change criticality
  - Implement priority inheritance (if you must…)

 © 2016 Gernot Heiser. Distributed under CC Attribution License