## Estimating Latency

Goal:
given information about events and their handlers ...
... estimate response times ...
$\ldots$. or devise scheduling to insure that deadlines are met.

## Definitions:

Latency [of an interrupt handler or dæmon task to an event]: Time from event occurrence to start of its handler or dæmon task.

Response time [of interrupt handler or dæmon task to an event]: Time from event occurrence to response. In class response assumed to be generated at completion of event's handler.

Let an event occur at $t_{1}$, its handler start at $t_{2}$ and finish at $t_{3} \ldots$
$\ldots$ then the latency is $t_{2}-t_{1}$ and response time $t_{3}-t_{1}$.

Run time [of interrupt handler or dæmon task]: The time needed to run ${ }^{1}$ on an unloaded system.
Actual run time [of interrupt handler or dæmon task]: The time needed to run in a particular situation (considering other CPU activity, etc.).
${ }^{1}$ For handlers (which run something like operating system subroutines), run time is total run time. For dæmons (which are tasks managed like any other tasks the OS is running) it's the amount of run time needed to generate a response. After the response the dæmon will wait for another event, in contrast to a handler which actually finishes execution. At a subsequent event a waiting dæmon moves from the wait state (if the event is handled by a dæmon), while a "fresh" call to a handler is made (if the event is handled state (if the eve
by a handler).

## Classes of problems:

- One-shot.

Assume that each event can occur at most once.

- Periodic exhaustive.

Events occur periodically, solution found exhaustively.

- Periodic statistical.

Events occur periodically, solution found using statistical methods.

## Archetypical Problems

Timing Estimation Problem:
Given:

- Events. When the events can occur
- Handlers and tasks for events. How long these will run.

Handlers and tasks generate a response by the end of their run.

- Scheduling algorithm, interrupt system, and related details.

Find:

- Worst-case response time for each event.


## Scheduling Problem:

Given:

- Events. When the events can occur.
- Deadlines for responses to these events.
- Handlers and tasks for events. How long these will run.
- Some details of interrupt hardware and scheduling.

Find:

- Scheduling priorities, algorithm, or other details so that deadlines are met.


## Trivial Example Problem

A RTS must react to a single event. Exactly $5 \mu \mathrm{~s}$ after the event occurs the NMI-request line will be asserted. (The delay is due to a slow sensor response.) The handler for this event requires $100 \mu \mathrm{~s}$ to generate a response to the event. The following are the timings of other system functions: context save, 100 ns ; task switch, $50 \mu \mathrm{~s}$; scheduler run time, $100 \mu \mathrm{~s}$; worst-case instruction completion, 50 ns ; masking and jump to address specified in IVT, 20 ns . Find the worst-case response time.

Solution:
The following occurs from event to response:
Total Time $/ \mu \mathrm{s}$ Time $/ \mu \mathrm{s} \quad$ Activity

| 0.00 | 0.00 | The event. |
| ---: | ---: | :--- |
| 5.00 | 5.00 | Time for sensor to generate request. |
| 5.05 | .05 | Instruction completion. |
| 5.07 | .02 | Mask and jump. |
| 5.17 | .10 | Save context. |
| 105.17 | 100.00 | Run handler. |

Note: In most problems the following times are ignored:
Context save, instruction completion, mask and jump.

## Pure Strong Priority, One-Shot Events

This covers two possible sets of conditions:
Interrupts:

- Responses are generated by interrupt handlers.
- There is no more than one event and handler for each interrupt level.
- Interrupts can only be masked by other interrupts.

There are no tasks or other unrelated activities which can mask interrupts.

- Only the run time of the interrupt handlers is significant.

Tasks:

- Responses are generated by tasks, with exactly one task per event.
- Priority scheduling is used with at most one task per priority level.
- The OS is task-preemptive.
- Only tasks specified in the problem need be considered.
(No other tasks will preempt or be selected before the tasks needed for responses.)
- The quantum is infinite

Interrupt Latency of Pure Strong Priority, One-Shot Events
Let
$\mathcal{E}$ be the set of event types
(e.g., $\mathcal{E}=\{$ button3, tempAlarm, eventX, $3, A\}$ ),
all events in $\mathcal{E}$ be one shot,
$t_{h}(e)$ be the run time of handler for $e \in \mathcal{E}$,
$p(e)$ be the strong priority level of the handler for $e \in \mathcal{E}$,
no two events have the same strong priority,
then worst case latency for event $e \in \mathcal{E}$ is

$$
\sum_{\substack{i \in \mathcal{E} \\ p(i)>p(e)}} t_{h}(i)
$$

In words: sum of run times of all higher priority interrupts.
Worst-case latency is encountered by $e$ if. . .
... at the same time as $e . .$.
...one or more higher priority events occur...
...and all other higher priority events occur before $e$ gets to run.

## Pure Weak Priority, One-Shot Events

This covers two possible sets of conditions:
Interrupts:

- Responses are generated by interrupt handlers.
$\diamond$ All interrupts share a single IRQ input.
$\diamond$ Interrupt to run determined by polling sequence.
- Interrupts only masked by other interrupts. (I.e., no tasks or other unrelated activities mask interrupts.)
- Only handler run time is significant.
E.g., context-switch and service-routine time can be ignored.

Tasks:

- Responses generated by tasks.
- Exactly one task per event.
- Priority scheduling used, at most one task per priority level.
$\diamond$ The OS is not task-preemptive.
- Only tasks specified in problem considered. (No other tasks will preempt or be selected before the tasks needed for responses.)
- The quantum is infinite.

Items that differ from the pure-strong-priority case are printed with slanted type and diamonds for bullets.

Pure Strong Priority, One-Shot Example Problem
Find the response time for all events in a system using pure-strongpriority interrupts and one-shot events. The events and interrupt handlers are described below.

| Event | Priority | Run Time |
| :---: | ---: | ---: |
| $A$ | 3 | $10 \mu \mathrm{~s}$ |
| $B$ | 7 | $15 \mu \mathrm{~s}$ |
| $C$ | 1 | $8 \mu \mathrm{~s}$ |

Solution:
Worst-case latency for $A$ : $t_{h}(B)=15 \mu \mathrm{~s}$.
Response time of $A$ is $t_{h}(B)+t_{h}(A)=25 \mu \mathrm{~s}$.
Worst-case latency for $B$ is 0 .
Response time of $B$ is $t_{h}(B)=15 \mu \mathrm{~s}$.
Worst-case latency for $C: t_{h}(A)+t_{h}(B)=25 \mu \mathrm{~s}$.
Response time of $C$ is $t_{h}(A)+t_{h}(B)+t_{h}(C)=33 \mu \mathrm{~s}$.
Note: $A$ and $C$ cannot simultaneously have worst-case response times.

Interrupt Latency of Pure Weak Priority, One-Shot Events Let
$\mathcal{E}$ be a set of one-shot event types
(e.g., $\mathcal{E}=\{$ button3, tempAlarm, eventX, $3, A\}$ ),
$t_{h}(e)$ be the run time of the handler for event $e \in \mathcal{E}$,
$p(e)$ be weak priority level of handler for $e \in \mathcal{E}$.
and no two events have the same weak priority,
then worst case latency for event $e \in \mathcal{E}$ is

$$
\left(\sum_{\substack{i \in \mathcal{E} \\ p(i)>p(e)}} t_{h}(i)\right)+\max _{\substack{i \in \mathcal{E}(e) \\ p(i)<p(e)}} t_{h}(i)
$$

In words:
Sum of run times of all higher-priority interrupts...
...plus longest run time of lower-priority interrupts.
Worst-case latency is encountered by $e$ if...
. . .just before $e \ldots$
...the lower-priority event with the longest handler occurs. . .
$\ldots$..and before this event's handler finishes...
...all events of higher priority than $e$ occur.

Pure Weak Priority, One-Shot Example Problem
Find the worst-case response time for all events in a system using pure-weak-priority interrupts and one-shot events. The events and interrupt handlers are described below.

| Event | Priority | Run Time |
| :---: | ---: | ---: |
| $A$ | 3 | $10 \mu \mathrm{~s}$ |
| $B$ | 7 | $15 \mu \mathrm{~s}$ |
| $C$ | 1 | $8 \mu \mathrm{~s}$ |

Solution:
Latency for $A: t_{h}(B)+t_{h}(C)=23 \mu \mathrm{~s}$.
Response time of $A$ is $t_{h}(B)+t_{h}(C)+t_{h}(A)=33 \mu \mathrm{~s}$.
Latency for $B$ is $t_{h}(A)=10 \mu \mathrm{~s}$.
Response time of $B$ is $t_{h}(A)+t_{h}(B)=25 \mu \mathrm{~s}$.
Latency for $C: t_{h}(A)+t_{h}(B)=25 \mu \mathrm{~s}$.
Response time of $C$ is $t_{h}(A)+t_{h}(B)+t_{h}(C)=33 \mu \mathrm{~s}$.

## Notes

$A$ and $C$ cannot simultaneously have worst-case response times.
Highest weak priority level does not guarantee lowest response time (This will be seen in an example below.)

Interrupt Latency of Strong and Weak Priority, One-Shot Events
Let
$\mathcal{E}$ be a set of event types
(e.g., $\mathcal{E}=\{$ button3, tempAlarm, eventX, $3, A\}$ ),
all events in $\mathcal{E}$ be one shot,
$t_{h}(e)$ be the run time of the handler for event $e \in \mathcal{E}$,
$p_{1}(e)$ be the strong priority level of the handler for event $e \in \mathcal{E}$,
$p_{2}(e)$ be weak priority level of handler for $e \in \mathcal{E}$.
Then worst-case latency for event $e \in \mathcal{E}$ is

$$
\sum_{\substack{i \in \mathcal{E} \\ p_{1}(i)>p_{1}(e)}} t_{h}(i)+\sum_{\substack{i \in \mathcal{E} \\ p_{1}(i)=p_{1}(e) \\ p_{2}(i)>p_{2}(e)}} t_{h}(i)+\max _{\substack{i \in \mathcal{E} \\ p_{1}(e)=p_{1}(e) \\ p_{2}(i)<p_{2}(e)}} t_{h}(i) .
$$

In words:
Sum of the run times of...
...all higher strong-priority and weak-priority handlers. . .
. . .plus longest run time of. . .
...the lower weak-priority handlers...
...that are at the same strong priority level as $e$.
This worst-case latency is encountered by $e$ if:
The longest-run-time handler at the same strong priority level...
...starts to run just before $e .$. .
. . .and all higher priority interrupts occur before $e$ 's handler starts.

Strong and Weak Priority, One-Shot Events
This covers two possible sets of conditions:
Interrupts:

- Responses are generated by interrupt handlers.
$\diamond$ An IRQ input can be used by any number of events.
$\diamond$ The interrupt to service determined by polling sequence... ...using a different sequence for each strong priority level.
- Interrupts can only be masked by other interrupts. There are no tasks or other unrelated activities which can mask interrupts.
- Only handler run time is significant.

Tasks:

- Responses are generated by tasks, exactly one task per event.
$\diamond$ Scheduling done in two rounds, both use priority scheduling.
$\diamond$ The first round is task-preemptive, the second round is non-taskpreemptive.
- Only tasks specified in the problem need be considered. (No other tasks will preempt or be selected before the tasks needed for responses.)
- The quantum is infinite.

Items that differ from the pure-weak-priority case are shown in slanted type with diamonds for bullets.

EE 4770 Lecture Transparency. Formatted 13:27, 23 December 1997 from Isli16.

Strong and Weak Priority, One-Shot Example Problem
Find the worst-case response time for all events in a system using strong and weak priority interrupts and one-shot events. The events and interrupt handlers are described below.

| Event | Strong <br> Priority | Weak <br> Priority | Run <br> Time |
| :---: | ---: | ---: | ---: |
| $A$ | 3 | 1 | $10 \mu \mathrm{~s}$ |
| $B$ | 2 | 3 | $15 \mu \mathrm{~s}$ |
| $C$ | 2 | 2 | $8 \mu \mathrm{~s}$ |
| $D$ | 2 | 1 | $50 \mu \mathrm{~s}$ |
| $E$ | 1 | 2 | $1 \mu \mathrm{~s}$ |
| $F$ | 1 | 1 | $2 \mu \mathrm{~s}$ |

Solution:
Event $A$
Event order: $A$ occurs any time.
Latency for $A: 0 \mu \mathrm{~s}$.
Response time of $A$ is $t_{h}(A)=10 \mu \mathrm{~s}$
Event $B$
Possible event order: $D, B, A$.
Latency for $B$ is $t_{l}(B)=t_{h}(A)+t_{h}(D)=60 \mu \mathrm{~s}$.
Response time of $B$ is $t_{l}(B)+t_{h}(B)=75 \mu \mathrm{~s}$.
Event $C$
Possible event order: $D, C, B, A$.
Latency for $C: t_{l}(C)=t_{h}(A)+t_{h}(B)+t_{h}(D)=75 \mu \mathrm{~s}$.
Response time of $C$ is $t_{l}(C)+t_{h}(C)=83 \mu \mathrm{~s}$.

Example, continued.

| Event | Strong <br> Priority | Weak <br> Priority | Run Time <br> Time |
| :---: | ---: | ---: | ---: |
| $A$ | 3 | 1 | $10 \mu \mathrm{~s}$ |
| $B$ | 2 | 3 | $15 \mu \mathrm{~s}$ |
| $C$ | 2 | 2 | $8 \mu \mathrm{~s}$ |
| $D$ | 2 | 1 | $50 \mu \mathrm{~s}$. |
| $E$ | 1 | 2 | $1 \mu \mathrm{~s}$. |
| $F$ | 1 | 1 | $2 \mu \mathrm{~s}$. |

Event $D$
Possible event order: $B, D, C, A$.
Latency for $D: t_{l}(D)=t_{h}(A)+t_{h}(B)+t_{h}(C)=33 \mu \mathrm{~s}$.
Response time of $D$ is $t_{l}(D)+t_{h}(D)=83 \mu \mathrm{~s}$.
Event $E$
Possible event order: $F, E, A, B, C, D$.
Latency for $E: t_{l}(E)=\sum_{e \in\{A, B, C, D, F\}} t_{h}(e)=85 \mu \mathrm{~s}$.
Response time of $E$ is $t_{l}(E)+t_{h}(E)=86 \mu \mathrm{~s}$.
Event $F$
Possible event order: $E, F, A, B, C, D$.
Latency for $F: t_{l}(F)=\sum_{e \in\{A, B, C, D, E\}} t_{h}(e)=84 \mu \mathrm{~s}$.
Response time of $F$ is $t_{l}(F)+t_{h}(F)=86 \mu \mathrm{~s}$.


```
* Time: (1)
terrupt A (1) requested
*)
    * Time: 1,007
    ** Time: 1,016
    F A (1) fimshed: lat. 0, dur. 10, resp. 10
    Handler for D (1) resumed
    Handler for D (1) finished: lat. 0., dur. 60., resp. 60.
    Handler for B (1) starting
    *)
    Handler for C (0) starting
    Handler for C (0) finished: lat. 75., dur. 8., resp. }83
    ================= D Worst-Case Latency and Run Time =======================
    Intrrupt B (2) requested.
    Time: 2,005
    * Time: 2006
    iterrupt A (2) requeste
    andler for A (2) starting.
    2,00
    * Time: 2,016
    Handler for B (2) resumed
    * Time: 2,030.
    Handler for C (1) starting.
    Handler for C (1) finished: lat. 23., dur. 8., resp. 31
    Handler for D (2) starting
    Handler for D (2) finished: lat. 33., dur. 50., resp. }8
    ============
    nterrupt F (0) requested
    * Tindler for F (0) starting
    terrupt E (0) reque
    terrupt A (3) requested.
    Hancer for ( (3) sreating.
```

    ** Time: \(\quad 3,007\)
    Interrupt B (3) requested
    ** Time: \(\quad 3,008\)
    Interrupt C (2) requested
    ** Time: \(\quad 3,009\)
    Interrupt D (3) requested
    Handler for A (3) finished: lat. 0, dur. 10, resp. 10
    Handler for \(\mathrm{F}(0)\) resumed.
    Handler for \(\mathrm{F}(0)\) preempted
    Handler for B (3) starting.
    ** Time: \(\quad 3,031\)
    Handler for B (3) finished: lat. 9, dur. 15, resp. 24
    Handler for \(\mathrm{F}(0)\) resumed.
    Handler for \(\mathrm{F}(0)\) preempted
    Handler for C (2) startin
    Handler for \(\mathrm{C}(2)\) finished: lat. 23, dur. 8 , resp. 31
    Handler for \(\mathrm{F}(0)\) resumed.
    Handler for \(\mathrm{F}(0)\) preempted
    Handler for D (3) starting.
    ** Time: 3,089
    Handler for D (3) finished: lat. 30, dur. 50, resp. 80
    Handler for \(\mathrm{F}(0)\) resumed.
    * Time: \(\quad 3,090\).
    Handler for \(\mathrm{F}(0)\) finished: lat. 0 ., dur. 85., resp. 85
    Handler for \(\mathrm{E}(0)\) starting.
    \({ }^{* *}\) Time: \(\quad 3,091\).
    ** Time: \(\quad 4,004\)
    ** Time: 4,005 .
    Interrupt E (1) requested.
Handler for $\mathrm{E}(1)$ starting
${ }^{*}$ * Time: $\quad 4,005$
${ }_{*}$ Interrupt F (1) requested.
** Time: $\quad 4,005$
Interrupt A (4) requested.
Handler for E (1) prempted.
Handler for E (1) preempted
${ }^{* *}$ Time: $\quad 4,007$
Interrupt B (4) requested
** Time: 4,008
Interrupt C (3) requested
** Time: 4,009
Interrupt D (4) requested.
${ }^{*}$ Time:
** Time: $\quad 4,015$
Handler for A (4) finished: lat. 0 , dur. 10 , resp. 10
Handler for E (1) resumed.
Handler for E (1) preempted.
${ }_{* *}^{\text {Hander Time: }} \quad 4,030$
Handler for B (4) finished: lat. 8, dur. 15, resp. 23

Handler for E (1) resumed
Handler for E (1) preempted.
Handler for C (3) star
** Time:
4,038
Handler for C (3) finished: lat. 22, dur. 8, resp. 30
Handler for E (1) resumed
Handler for E (1) preempted.
${ }_{* *}^{\text {Handler for }}$ D (4) starting.
** Time: 4,088
Handler for D (4) finished: lat. 29, dur. 50, resp. 79
Handler for E (1) resimed
Handler for E (1) resumed
** Time: 4,089.
Handler for E (1) finished: lat. 0., dur. 84., resp. 84.
Handler for F (1) starting.
Handler for $\mathrm{F}(1)$ finished: lat. 84., dur. 2., resp. 86. Simulation completed.

## Perturbations Example Problem

Find the worst-case response time for all events in a system using strong and weak priority interrupts and one-shot events. The events and interrupt handlers are described below. The handler must be run once for each event occurrence, even if an event occurs a second time before the first the handler was run for the first occurrence.

| Event | Strong <br> Priority | Weak <br> Priority | Run <br> Time | Event <br> Occurrences |
| :---: | :---: | :---: | :---: | :--- |
| $A$ | 3 | 1 | $10 \mu \mathrm{~s}$ | Occurs once, any time. |
| $B$ | 2 | 3 | $15 \mu \mathrm{~s}$ | Occurs twice, any times. |
| $C$ | 2 | 2 | $8 \mu \mathrm{~s}$ | Occurs once, 45 to $50 \mu \mathrm{~s}$ after event $A$. |
| $D$ | 2 | 1 | $50 \mu \mathrm{~s}$ | Occurs once, any time. |
| $E$ | 1 | 2 | $1 \mu \mathrm{~s}$ | Occurs three times, with $>100 \mu \mathrm{~s}$ separation. |
| $F$ | 1 | 1 | $2 \mu \mathrm{~s}$ | Occurs once. |

Solution:
Event $A$
Event order: $A$ can occur any time.
Latency for $A: 0 \mu \mathrm{~s}$.
Response time of $A$ is $t_{h}(A)=10 \mu \mathrm{~s}$.
Event $B$
Since event-type $B$ can occur twice...
...an event $B$ might have to wait...
$\ldots$ for a previous occurrence of $B$ to be handled.
Possible event order: $D, B_{1}, B_{2}, A$.
Latency for $B_{2}$ is $t_{l}(B)=t_{h}(A)+t_{h}(B)+t_{h}(D)=75 \mu \mathrm{~s}$.
Response time of $B_{2}$ is $t_{l}(B)+t_{h}(B)=90 \mu \mathrm{~s}$.

Example, continued.

| Event | Strong <br> Priority | Weak <br> Priority | Run <br> Time | Event <br> Occurrences |
| :--- | :---: | :---: | ---: | :--- |
| $A$ | 3 | 1 | $10 \mu \mathrm{~s}$ | Occurs once, any time. |
| $B$ | 2 | 3 | $15 \mu \mathrm{~s}$ | Occurs twice, any times. |
| $C$ | 2 | 2 | $8 \mu \mathrm{~s}$ | Occurs once, 45 to $50 \mu$ s after event $A$. |
| $D$ | 2 | 1 | $50 \mu \mathrm{~s}$ | Occurs once, any time. |
| $E$ | 1 | 2 | $1 \mu \mathrm{~s}$ | Occurs three times, with $>100 \mu$ s separation. |
| $F$ | 1 | 1 | $2 \mu \mathrm{~s}$ | Occurs once. |
|  |  |  |  |  |
|  | Event $C$ |  |  |  |

$C$ only occurs after event $A$ so:
Possible event order: $D, C, B_{1}, B_{2}$.
Latency for $C$ : $t_{l}(C)=2 t_{h}(B)+t_{h}(D)=80 \mu \mathrm{~s}$.
Response time of $C$ is $t_{l}(C)+t_{h}(C)=88 \mu \mathrm{~s}$.
Event $D$.
$C$ and $A$ cannot both occur within $D$ 's latency period.
Possible event order: $B_{1}, D, A, B_{2}$
Latency for $D: t_{l}(D)=t_{h}(A)+2 t_{h}(B)=40 \mu \mathrm{~s}$.
Response time of $D$ is $t_{l}(D)+t_{h}(D)=90 \mu \mathrm{~s}$.

## Example Problem Solution Details



```
Interrupt B (4) requested.
Handler for B (4) starting
** Time: 2,005
Interrupt D (2) requested
** Time: 2,006
Handler for B (4) preempted
Handler for B (4) preempted
*** Time: 
Interrupt B (5) requested
** Time: 2,016
Handler for A (1) finished: lat. 0, dur. 10, resp. }1
Handler for B (4) resumed.
** Time: 2,030.
Handler for B (4) finished: lat. 0., dur. 25., resp. 25
Handler for B (5) starting.
** Time: 
Handler for B (5) finished: lat. 23., dur. 15., resp. 38.
Handler for D (2) starting.
** Time: 2,095.
Handler for D (2) finished: lat. 40., dur. 50., resp. 90.
** Time: 3,004
** Time: 3,005.
Interrupt F (0) requested.
Handler for F (0) starting.
** Time: }3,00
*** Time: 
** Time: 3,006
Handler for F (0) preequest.
Handler for F (0) preempte
** Time: 3,007
*** Trrupt B (6) requested.
** Time: 3,008
*** Inrupt B (7) requested
** Time: 3,009
Interrupt D (3) requested
Time: 3,016
Handler for A (2) finished: lat. 0, dur. 10, resp. }1
Handler for F (0) resumed.
Handler for F (0) preempted
Handler for B (6) starting.
** Time: 3,031
Handler for B (6) finished: lat. 9, dur. 15, resp. 24
Handler for F (0) resumed.
Handler for F (0) preempted
Handler for B (7) starting.
** Time: 3,046
Handler for B (7) finished: lat. 23, dur. 15, resp. }3
Handler for B (7) finished: lat
Handler for F (0) preempted.
Handler for D (3) starting.
```



Relative Timing
Since time of occurrences arbitrary...
...in a system with several periodic event types,...
...the relative timing of the events is arbitrary.
Relative Timing Example
Suppose event-type $A$ is periodic with $t_{b}(A)=10 \mathrm{~ms} \ldots$
$\ldots$ and $B$ is periodic also with $t_{b}(B)=10 \mathrm{~ms}$.
Then $A$ and $B$ could occur at:..
...the same time,...
$\ldots$ or $A$ could follow $B$ by any time $<10 \mathrm{~ms}$.
But, because their periods are the same,...
...the time from $A$ to $B$ is fixed,..
...even though this time is not known in advance.

## Latency Estimation with Periodic Events

Exhaustive Method
By trial-and-error, find the worst-case scenario.
Statistical Method
Find average effect of relatively short handlers.

## 16-35

## Periodic-Interrupts Example Problem

A RTS has three event types, $A, B$, and $C$. All event types are periodic; their periods, the run time of their handlers, and their priority levels appear in the table below. Find the latency and response time for each event type.

| Event <br> Name | Strong <br> Priority | Period <br> $t_{b} / \mu \mathrm{s}$ | Handler Run Time <br> $t_{h} / \mu \mathrm{s}$ |
| :---: | :--- | ---: | ---: |
| $A$ | 3 | 23 | 5 |
| $B$ | 2 | 100 | 20 |
| $C$ | 1 | 36 | 2 |

Solution:
Latency for $A: 0 \mu \mathrm{~s}$.
Response time of $A$ is $t_{r}(A)=t_{h}(A)=5 \mu \mathrm{~s}$.
Latency for $B$ : $t_{l}(B)=t_{h}(A)=5 \mu \mathrm{~s}$.
Response time of $B$ is $t_{r}(B)=2 t_{h}(A)+t_{h}(B)=30 \mu \mathrm{~s}$.
( $B$ occurs just after $A$; before $B$ finishes $A$ occurs a second time.)
Latency for $C: t_{l}(C)=2 t_{h}(A)+t_{h}(B)=30 \mu \mathrm{~s}$.
Response time of $C$ is $t_{r}(C)=2 t_{h}(A)+t_{h}(B)+t_{h}(C)=32 \mu$.

## Statistical Method of Latency Estimation

Motivation: exhaustive method may be too time consuming. Statistical Method Idea

Consider average - not exact-number of times...
...one handler interrupts another.
Illustration
Consider two periodic events, $A$ and $B$.
Let interrupt $A$ have higher strong priority than $B$.
Let $t_{h}(A)=100 \mu \mathrm{~s}, t_{b}(A)=300 \mu \mathrm{~s}, t_{h}(B)=450 \mu \mathrm{~s}$, and $t_{b}(B)=10 \mathrm{~s}$.
Exhaustive Method on Illustration
$B$ can be interrupted by $A 2$ or 3 times.
WC run-time for $B$ then $t_{a}(B)=t_{h}(B)+n t_{h}(A)$,
where $n$ is number of times $B$ interrupted:
Substituting yields $t_{a}(B)=750 \mu \mathrm{~s}$.

WC Run Using Average Number of Interruptions
Ignoring fact that $n$ is an integer...
$\ldots$.let $n=t_{a}(B) / t_{b}(A)$.
This average contains unknown value, $t_{a}(B)$.
Solving for $t_{a}(B)$ yields $t_{a}(B)=\frac{t_{h}(B)}{1-\frac{t_{h}(A)}{t_{b}(A)}}$.
Average Method on Illustration
Substituting yields $t_{a}(B)=675 \mu \mathrm{~s}$.
Difference, $750 \mu \mathrm{~s}-675 \mu \mathrm{~s}=75 \mu \mathrm{~s}$, too big to ignore...
...but as $n$ increases the difference drops.

First definitions, then details of the statistical method.

Load Set
Let $Y$ denote an event.
The load set for $Y$ is ...
... the set of events that load $Y$.

## Loading Factor

Let $Y$ denote an event..
$\ldots$ and $\mathcal{X}$ be the load set for $Y$.
Then the loading factor for $Y$ 's handler is

$$
l_{f}(Y)=1-\sum_{e \in \mathcal{X}} \frac{t_{h}(e)}{t_{b}(e)}
$$

In words: $Y$ 's loading factor is fraction of CPU time ...
... available to $Y$ 's handler ...
$\ldots$ after accounting for $\mathcal{X}$, the load set.

## Loaded Duration

Let $Y$ denote an event with loading factor $l_{f}(Y)$.
Then the loaded duration of event $Y$ 's handler is

$$
t_{h}^{\prime}(Y)=\frac{t_{h}(Y)}{l_{f}(Y)}
$$

Note: the actual duration, $t_{a}(Y)$, can be longer.

## General Solution Technique

Start at the highest strong-priority level.
For all events in a strong-priority level:

- Find the loading factor, compute the loaded duration.
- Exhaustively compute the actual duration (consider only lowerpriority events which do not load the event being considered).
- Exhaustively compute the latency.
- Exhaustively compute the response time.

Repeat for the next lower strong-priority level or finish if at the lowest strong priority.

Statistical-Latency-Estimation Example Problem
A RTS has three event types, all periodic. Their period, priority, and the run time of their handlers is listed in the table below. Find the latency and response time for each event.

| Event <br> Name | Strong <br> Priority | Period <br> $t_{b}$ | Handler Run <br> Time, $t_{h}$ |
| :---: | :--- | ---: | :---: |
| $A$ | 3 | $10 \mu \mathrm{~s}$ | $0.5 \mu \mathrm{~s}$ |
| $B$ | 2 | 50 ms | $37 \mu \mathrm{~s}$ |
| $C$ | 1 | 1 s | $40 \mu \mathrm{~s}$ |

Solution:

- Event $A$

Latency, $t_{l}(A)=0 \mu \mathrm{~s}$.
Response time, $t_{r}(A)=t_{l}(A)+t_{h}(A)=0.5 \mu \mathrm{~s}$.

- Event B
$\frac{t_{h}(B)}{t_{h}(A)}=74>\theta_{l}=50$, therefore $A$ loads $B$.
Loading factor for $B$ is $l_{f}(B)=1-\frac{t_{h}(A)}{t_{b}(A)}=0.95$.
Loaded duration: $t_{h}^{\prime}(B)=\frac{t_{h}(B)}{l_{f}(B)}=38.95 \mu \mathrm{~s}$.
Latency, $t_{l}(B)=t_{h}(A)=0.5 \mu \mathrm{~s}$.
Response time, $t_{r}(B)=t_{l}(B)+t_{h}^{\prime}(B)=39.45 \mu \mathrm{~s}$.


## Perturbations

## Quasi-Periodic Events

Occur regularly, but not with fixed period.
Examples:

- Period ranging from 100 ms to 80 ms .
- No more than 5 times in 10 ms interval.
- At $t=0$ and exactly $27 \mu \mathrm{~s}$ after previous occurrence handled.


## Load of Handlers for Quasi-Periodic Events

Can sometimes find a worst-case $t_{b}$.
This would be used in load factors.
The method for determining the WC $t_{b}$ depends upon details of quasiperiodic event.


Example, continued.

| Event | Strong <br> Name | Weak <br> Pri. | Pri. | Handler <br> Run Time |
| :---: | :--- | :--- | ---: | :--- | Occurrence $\quad$.

Event $E$
Loaded by $A, B$, and $D$ :

$$
l_{f}(E)=1-\frac{t_{h}(A)}{t_{b}(A)}-\frac{t_{h}(B)}{t_{b}(B)}-\frac{t_{h}(D)}{t_{b}(D)}=0.717 .
$$

Worst-case latency includes
...time for $C \ldots$
$\ldots$...plus time for $D$ 's waiting after $C$ finishes...
$\ldots$ plus 3 A's and $2 B$ 's:
$t_{l}(E)=3 t_{h}(D)+3 t_{h}(A)+2 t_{h}(B)+t_{a}(C)=979.33 \mu \mathrm{~s}$.
Actual duration is similar, except $E$ occurs before $C$.
$t_{a}(E)=\frac{t_{h}(E)}{l_{f}(E)}+t_{a}(C)+3 t_{a}(D)+3 t_{a}(A)+2 t_{a}(B)=2.374 \mathrm{~ms}$.
Response time must not count $C$ twice:
$t_{r}(E)=t_{l}(E)+\frac{t_{h}(E)}{l_{f}(E)}=2.374 \mathrm{~ms}$.

| Event <br> Name | Strong <br> Pri. | Weak <br> Pri. | Handler <br> Run Time | Occurrence |
| :---: | :--- | :--- | ---: | :--- |
| $A$ | 4 | 2 | $3 \mu \mathrm{~s}$ | Periodic, $t_{b}(A)=20 \mu \mathrm{~s}$. |
| $B$ | 4 | 1 | $2 \mu \mathrm{~s}$ | From $7 \mu \mathrm{~s}$ to $13 \mu \mathrm{~s}$ after event $A$, if at all. |
| $C$ | 3 | 1 | $700 \mu \mathrm{~s}$ | Periodic, $t_{b}(C)=27 \mathrm{~ms}$. |
| $D$ | 3 | 2 | $11 \mu \mathrm{~s}$ | No more than 3 times in any 1 ms interval. |
| $D$ |  |  |  | At most 2 unresponded events held. |
| $E$ | 2 | 1 | 1 ms | Periodic, $t_{b}(E)=100 \mathrm{~ms}$. |
| $F$ | 1 | 1 | 500 ms | Anytime after resp. to prev. occur. |

Event $F$

$$
\text { All other events load } F \text { : }
$$

$$
l_{f}(F)=1-\sum_{e \in\{A, B, C, D, E\}} \frac{t_{a}(e)}{t_{b}(e)}=0.681
$$

Duration

$$
t_{a}(F)=\frac{t_{h}(F)}{l_{f}(F)}=734.1 \mathrm{~ms}
$$

Latency of $F$ is the same as response time of $E$ since there is only one interrupt at strong levels 1 and 2 and since neither interrupt will recur until after handler finishes.

$$
\begin{aligned}
& t_{l}(F)=t_{r}(E)=2.374 \mathrm{~ms} \\
& t_{r}(F)=t_{l}(F)+t_{a}(F)=736.5 \mathrm{~ms}
\end{aligned}
$$

Because of event $F$, worst-case load is 1 .

## Example Problem Solution Details

Simulator Output Format

> Output shows worst-case scenarios for some events.

As with simulator output appearing above, latency and response times shown are only worst case for a few events, see title at the head of each run.
Events in a handler's load set are not shown while a handler is running. For example, $C$ is loaded by $A$ and $B$, so while $C$ is running occurrences of $A$ and $B$ are not shown.
Just before a handler starts a message is printed for each active event in its load set, indicating that those events will be ignored (and the run time of the handler is computed using the loading
factor), for example, "Event A simulated using handler load." If the event is already being ignored no message is printed. Similarly, before a handler starts a message is printed for each ignored no message is printed. Similarly, before a handler starts a message is printed for each
event which will no longer be ignored. For example, "Normal simulation of A resuming." event which will no longer be ignored. For example, "Normal simulation of A resuming.
Sometimes both messages are printed (the software is not yet polished), use the second message.

- Loaded Event Handlers

Event C: ld. factor, 0.75; ld. dur, 933.333 load set A, B
Event E: Id. factor, 0.75; ld. dur, 1333.33 load set A, B
${ }_{* *}$ Event F: ld. factor, 0.714074; ld. dur, 700207. load set A, B, C, E
** Starting simulation...
$=================\mathrm{C}(0)$ Worst-Case Latency and Response. $======================$
** Time: 1,000 .
** Time: $\quad 1,000$.
Interrupt D $(0)$ requested.
Interrupt D ( 0 ) requested.
Handler for D $(0)$ starting, time remaining 11
$\begin{array}{cc}* * \text { Time: } & 1,000 . \\ \text { Interrupt A } & 0) \text { reque }\end{array}$
Handler for D $(0)$ preempted.
Handler for $\mathrm{A}(0)$ starting, time remaining 3
$* *$
** Time: $\quad 1,000$
Interrupt C ( 0 ) requested.
** Time:
1,000
** Time: 1,000 .
Interrupt D (1) requested.
** Time:
$1,003$.
1,003.
Handler for A (0) finished:
Handler for $\mathrm{D}(0)$ resumed time $3 .=$ response time 3 .
${ }_{* *}^{\text {Handler for }}$ D ( 0 ) resumed, time remaining 11..
Interrupt B (0) requeste
Handler for D ( 0 ) preempted.
Handler for B (0) starting, time remaining 2
$\begin{array}{ll}\text { ** Time: } & 1,009 \\ \text { Handler for }\end{array}$
Handler for B ( 0 ) finished:
Hatency $0+$ duration $2=$ response time 2
Handler for D (0) resumed, time remaining $7 .$.
${ }^{* *}$ Time: $\quad 1,016$.
** Time: $\quad 1,016$.
Handler for D $(0)$
latency $0 .+$ duration $16 .=$ response time 16 .

Handler for
${ }^{*}$ D (1me:
1,020 . starting, time remaining
11
** Time: $\quad 1,020$.
Interrupt A (1) requested.
Hand
Handler for $\mathrm{A}(1)$ starting, time remaining 3
${ }_{*}, 023$.
Handler for $\mathrm{A}(1)$ finishe
latency $0 .+$ duration $3 .=$ response time 3.
${ }_{* * *}$ Timer for $\mathrm{D}(1)$ resumed, time remaining $7 .$.
** Time: 1,025
${ }_{* *}$ Interrupt $\mathrm{D}(2)$ requested.
** Time: 1,027
Interrupt B (1) requested.
Handler for D (1) preempted
${ }_{* *}^{\text {Handler for }} \mathrm{B}(1)$ starting, time remaining 2
Handler for B ${ }^{1,029}$
Hander latency $0+$ duration $2=$ response time 2
Handler for D (1) resumed, time remaining 3..
** Time: $\quad 1,032$.
Handler for $\mathrm{D}(1)$ finished:
latency $16 .+$ duration 16. = response time 32
Handler for $\mathrm{D}(2)$ starting, time remaining 11
** Time: $\quad 1,040$.
Interrupt A (2) requested.
Handler for D (2) preempted.
Handler for A (2) starting, time remaining 3
** Time: 1,043 .
Handler for A (2) finished:
Cncy $0 .+$ duration $3 .=$ response time 3 .
** Time: $\quad 1,046$.
Handler for D (2) finished:
$\begin{aligned} \text { latency 7. }+ \text { duration } 14 .\end{aligned}=$ response time 21.
Event A simulated using handler load
Event B simulated using handler load.
Handler for C (0) starting, time remaining 700
** Time: $1,979.33$
Handler for $\mathrm{C}(0)$ finished:
latency $46 .+$ duration $933.333=$ response time 979.333

Loaded Event Handler
Event C: ld. factor, 0.75; ld. dur, 933.333 load set A, B
Event F: Id. factor, 0.714074 ; ld. dur, 700207. load set A, B, C, E
Event A simulated using handler load.
Event B simulated using handler load.
** Starting simulation...

* Time: 1,000 .

Interrupt C ( 0 ) requested.
Handler for C (0) starting, time remaining 700

* Time: $\quad 1,000$
** Time: 1,000
Interrupt D (1) requ
** Time: $1,933.33$
Handler for C (0) finished:
latency $0 .+$ duration $933.333=$ response time 933.333
Normal simulation of A resuming
Normal simulation of B resuming
Handler for $\mathrm{D}(0)$ starting, time remaining 11
** Time: 1,934 .
Interrupt B ( 0 ) requested.
preempted
andler for B (0) starting, time remaining 2
Time: ${ }^{1,936 .}$
latency $0 .+$ duration $2 .=$ response time 2
Handler for D ( 0 ) resumed, time remaining 10.3333.
** Time: 1,941.
Interrupt A (0) requested.
Handler for D ( 0 ) preempted.
Handler for A (0) starting, time remaining 3
* Time: 1,944.

Handler for A (0) finished:
ander for ( 0 ) duration 3. $=$ response time 3
Fandler for $\mathrm{D}(0)$ resumed, time remaining 5.33333
Handler for D ( 0 ) finish
latency $933.333+$ duration $16 .=$ response time 949.333
Event A simulated using handler load.
Event B simulated using handler load.
Normal simulation of A resuming
Normal simulation of B resuming
Handler for D (1) starting, time remaining 11

* Time: $1,954$.

Handler for D (1) preempted.
Handler for B (1) starting, time remaining 2
** Time: $\quad 1,956$.
Handler for B (1) finished:
Handler for D (1) resumed, time remaining 6.33333 .

Loaded Event Handlers
vent C: Id. factor, 0.75 ; ld. dur, 933.333 load set A, B
Event E: ld. factor, 0.717; ld. dur, 1394.7 load set A, B, D
Event F: ld. factor, 0.681074; ld. dur, 734135. load set A, B, C, D, E
Event A simulated using handler load.
Event D simulated using handler load.
** Time: $\quad 1,000$.
interrupt C ( 0 ) requested.
Normal simulation of D resuming
Handler for C (0) starting, time remaining 700

* Time: 1,000
** Time: F (0) requeste
Time: 1,000
* Time: $1,288.33$
nterrupt D (0) requ
** Time: $1,621.67$
Interrupt D (1) reques
** Time: $1,933.33$
Handler for $\mathrm{C}(0)$ finished:
latency 0 . + duration $933.333=$ response time 933.333
vent D simulated using handler load.
Normal simulation of A resuming
Normal simulation of B resuming
Handler for D ( 0 ) starting, time remaining 11
** Time: $\quad 1,934$.
nterrupt A (0) requested.
Handler for D (0) preempted.
Handler for A (0) starting, time remaining 3
* Time: $1,937$.

Handler for A (0) finished:
Latency $0 .+$ duration 3. $=$ response time 3
Handler for D ( 0 ) resumed, time remaining 10.3333

* Time: 1,941.

Handler for D ( 0 ) preempt
ander for D (0) preempted.
starting, time remaining 2

* Time: 1,943.
andler for $\mathrm{B}(0)$ finished:
latency $0 .+$ duration $2 .=$ response time 2.
fandler for $\mathrm{D}(0)$ resumed, time remaining 6.33333 .
* Time: $1,949.33$

Handler for D (0) finished:
latency $645 .+$ duration $16 .=$ response time 661
vent A simulated using handler load
Event B simulated using handler load.
Normal simulation of A resuming.
Normal simulation of B resuming
Normal simulation of D resuming.
Handler for D (1) starting, time remaining 11
** Time: $\quad 1,954$.
Interrupt A (1) requested.
Handler for A (1) starting, time remaining 3
** Time: 1,955 .
Interrupt D (2) requested
Handler for A (1) finished
latency $0 .+$ duration $3 .=$ response time 3
${ }_{* *}$ Handler for $\mathrm{D}(1)$ resumed, time remaining 6.33333 .
** Time: $\quad 1,961$.
Interrupt B (1) requested.
Handler for D (1) preempted
Handler for B (1) starting, time remaining 2
** Time: 1,963.
Handler for B (1) finished
${ }_{*}^{\text {H* Time: }} 1,965.33$
Handler for D (1) finished
latency $327.667+$ duration 16. $=$ response time 343.667
Event A simulated using handler load
Event B simulated using handler loa
Normal simulation of $A$ resumi
Normal simulation of B resuming.
Normal simulation of D resuming.
Handler for D (2) starting, time remaining 11
** Time: 1,974.
Interrupt A (2) requested.
Handler for D (2) preempted
Handler for A (2) starting, time remaining 3
** Time: $1,977$.
Handler for A (2) finished:
atency $0 .+$ duration $3 .=$ response time 3
Handler for D (2) resumed, time remaining 2.33333
** Time: 1,979.33
Handler for D (2) finished
atency 10.33s3 \& duration 14. $=$ response time 24.3333
Event B simulated using handler load
Event B simulated using handler load
Event D simulated using handler load
Handler for $\mathrm{E}(0)$ starting, time remaining 1000
Hadler for $\mathrm{E}(0)$ fin
Handler for $\mathrm{E}(0)$ finished
latency 979.333 + duration $1394.7=$ response time 2374.03
Event C simulated using handler load
Handler for F ( 0 ) starting, time remaining 500000
** Time: 737,509.
Handler for $\mathrm{F}(0)$ finished
latency $2374.03+$ duration 734135. $=$ response time 73650.

