

EXECUTION 47 · SIMULATED

14:03:19 AGENT KILL SWITCH

DISABLED

14:03:20 new execution

REJECTED

14:03:34 email.send(customer)

SENT

EXECUTION 47 · SIMULATED

14:03:19 AGENT KILL SWITCH

DISABLED

14:03:20 new execution

REJECTED

14:03:34 email.send(customer)

SENT

Your AI Kill Switch Is Probably **Broken**

Feature flags, cancellation semantics, and proving an agent actually stopped

FIRE DRILL #1 · SIMULATED

Disabled, but still running

EXECUTION 47 · RUNNING

[14:03:12]	[agent-core]	EXECUTION 47	STARTED
[14:03:15]	[tool:crm]	lookup_customer	COMPLETE
[14:03:17]	[tool:jira]	create_ticket	COMPLETE
[14:03:18]	[queue:outbound]	enqueue reconcile-job	ACCEPTED

Simulated: local runtime · mocked queue and worker · mocked CRM, Jira and SMTP

FIRE DRILL #1 · SIMULATED

Disabled, but still running

EXECUTION 47 · RUNNING

[14:03:12]	[agent-core]	EXECUTION 47	STARTED
[14:03:15]	[tool:crm]	lookup_customer	COMPLETE
[14:03:17]	[tool:jira]	create_ticket	COMPLETE
[14:03:18]	[queue:outbound]	enqueue reconcile-job	ACCEPTED
[14:03:19]	[control-plane]	AGENT KILL SWITCH	DISABLED
[14:03:20]	[agent-core]	new execution	REJECTED

Simulated: local runtime · mocked queue and worker · mocked CRM, Jira and SMTP

FIRE DRILL #1 · SIMULATED

Disabled, but still running

EXECUTION 47 · RUNNING

[14:03:12]	[agent-core]	EXECUTION 47	STARTED
[14:03:15]	[tool:crm]	lookup_customer	COMPLETE
[14:03:17]	[tool:jira]	create_ticket	COMPLETE
[14:03:18]	[queue:outbound]	enqueue reconcile-job	ACCEPTED
[14:03:19]	[control-plane]	AGENT KILL SWITCH	DISABLED
[14:03:20]	[agent-core]	new execution	REJECTED
[14:03:34]	[worker:async]	email.send(customer)	EXECUTING
[14:03:34]	[external:smtp]	message payload	SENT

DISABLED FOR

00:15

Still executing.

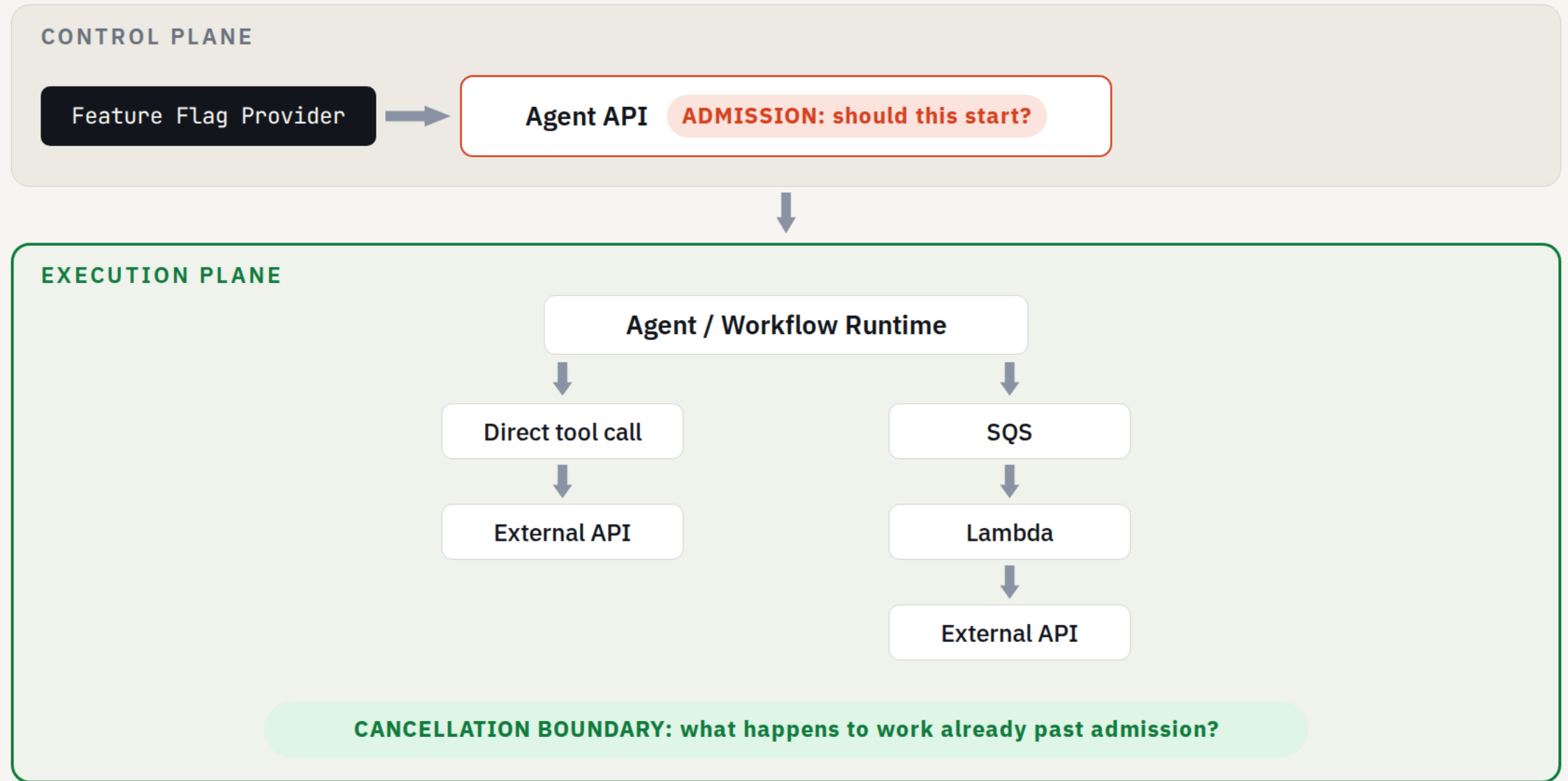
Irreversibly.

The switch worked.

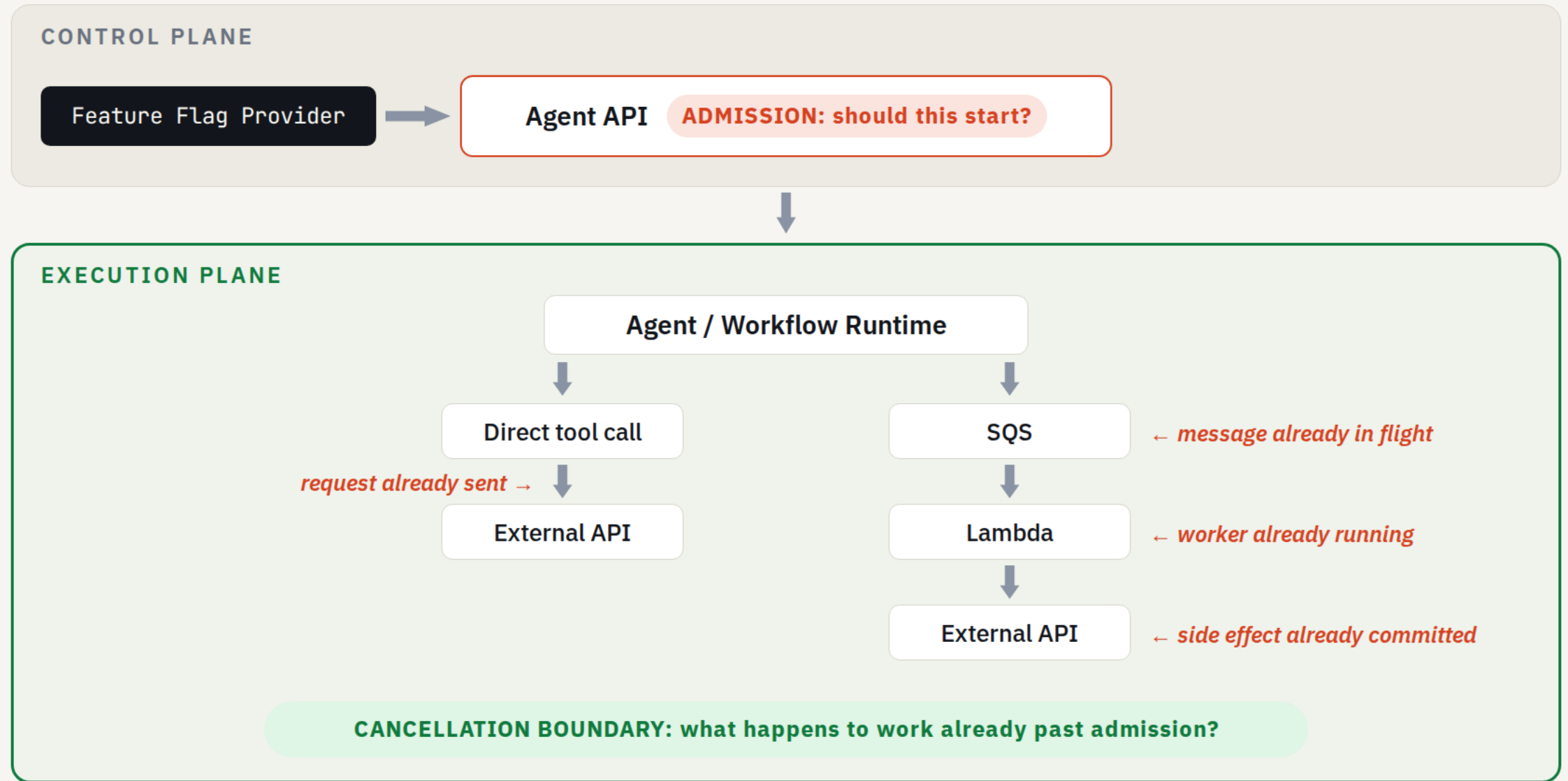
The system did not stop.

admission $\neq$ cancellation

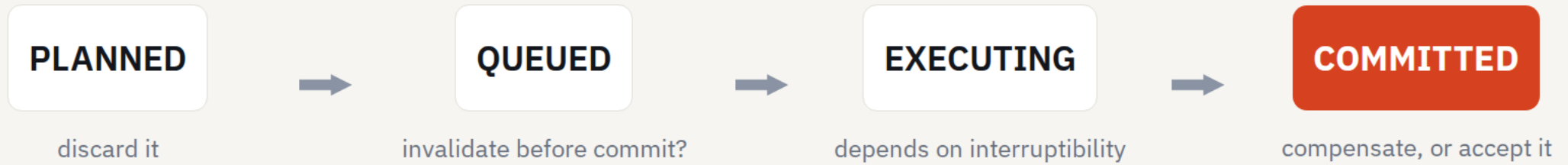
GHOSTS ALREADY IN MOTION



GHOSTS ALREADY IN MOTION



THE POINT OF NO RETURN



Cancellation prevents future work. Compensation deals with work cancellation arrived too late to prevent.

THIS ISN'T HYPOTHETICAL

2012

A repurposed flag triggered obsolete trading code left on one server.

~45 minutes · ~\$460M

while engineers tried to make it stop. Not an AI agent, but a pre-AI reminder that control intent and runtime behavior can diverge catastrophically.

Knight Capital · SEC Release No. 34-70694

THE KILL-SWITCH CONTRACT · FIVE RESPONSIBILITIES

1 **ADMISSION**

prevent new autonomous work

THE KILL-SWITCH CONTRACT · FIVE RESPONSIBILITIES

1 **ADMISSION** prevent new autonomous work

2 **PROPAGATION** tell active executions cancellation was requested

THE KILL-SWITCH CONTRACT · FIVE RESPONSIBILITIES

1 **ADMISSION**

prevent new autonomous work

2 **PROPAGATION**

tell active executions cancellation was requested

3 **INTERRUPTION**

stop what can still safely be stopped

THE KILL-SWITCH CONTRACT · FIVE RESPONSIBILITIES

1 **ADMISSION**

prevent new autonomous work

2 **PROPAGATION**

tell active executions cancellation was requested

3 **INTERRUPTION**

stop what can still safely be stopped

4 **COMPENSATION**

reconcile effects that already happened

THE KILL-SWITCH CONTRACT · FIVE RESPONSIBILITIES

- | | | |
|---|---------------------|--|
| 1 | ADMISSION | prevent new autonomous work |
| 2 | PROPAGATION | tell active executions cancellation was requested |
| 3 | INTERRUPTION | stop what can still safely be stopped |
| 4 | COMPENSATION | reconcile effects that already happened |
| 5 | VERIFICATION | prove autonomous activity ceased, within what we can observe |

ADMISSION + PROPAGATION

CONTROL PLANE

FLAG = OFF

↓ does not imply ↓

EXECUTION STOPPED?

EXECUTION PLANE

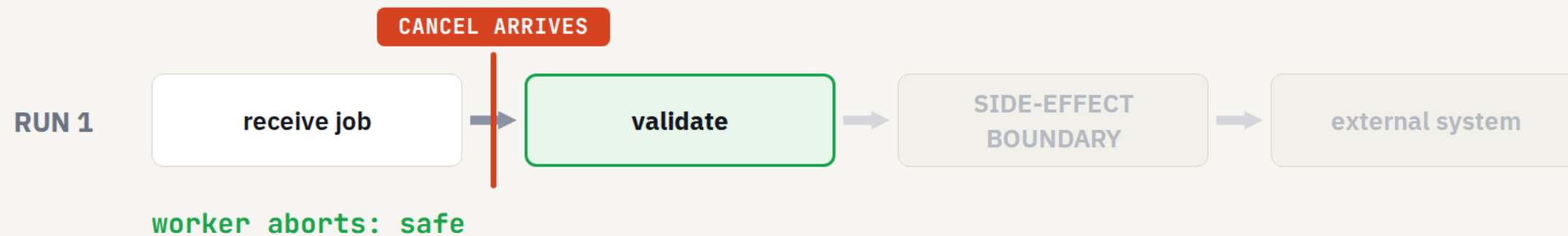
execution 62

generation 18 → 19**CANCEL_REQUESTED**

stale generation = reject stale work

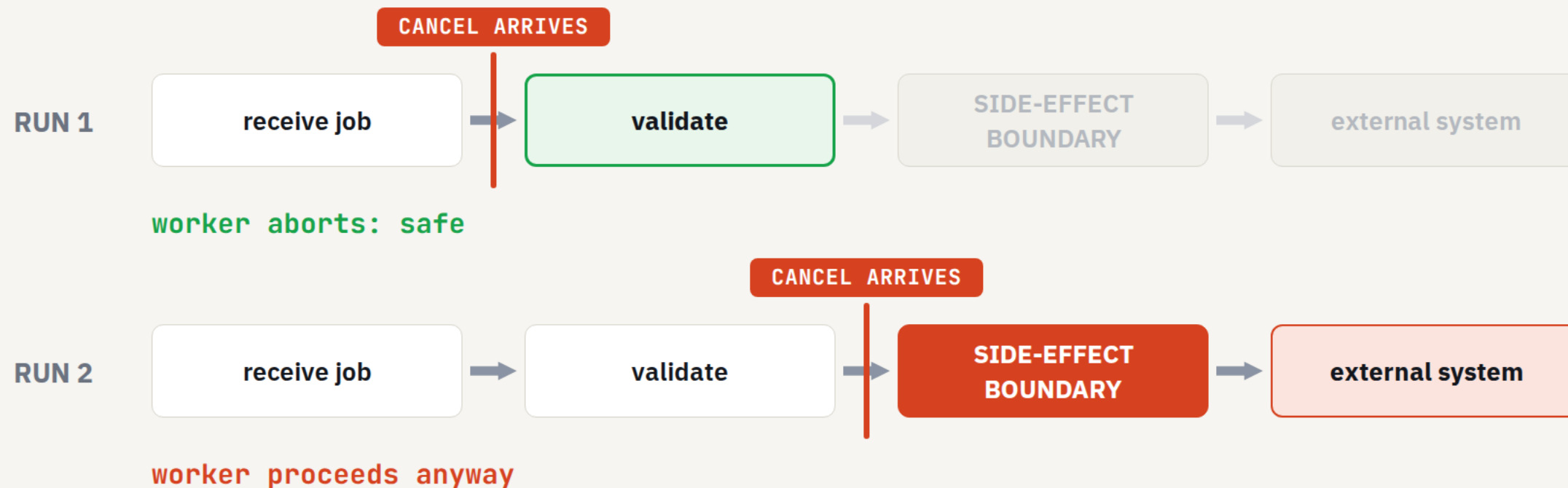
INTERRUPTION IS THE HARD PART

EXECUTION PLANE · WORKER



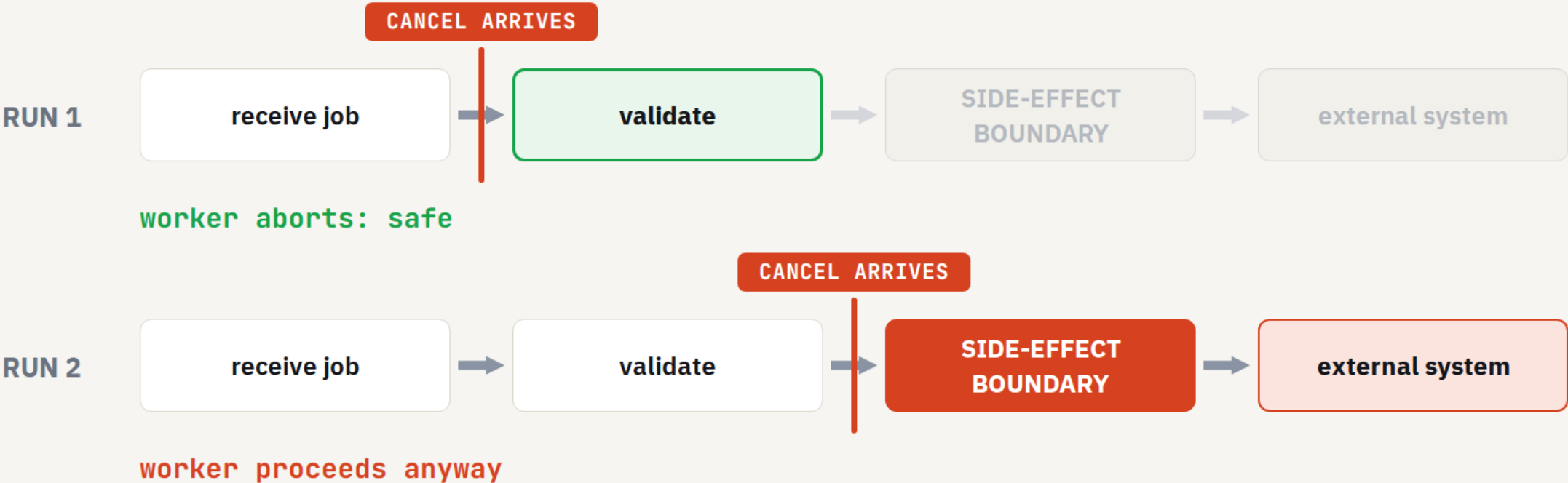
INTERRUPTION IS THE HARD PART

EXECUTION PLANE · WORKER



INTERRUPTION IS THE HARD PART

EXECUTION PLANE · WORKER



valid at check $\neq$ guaranteed valid at commit

YOU CANNOT CANCEL THE PAST

REVERSIBLE

Draft created

Delete it. A genuine inverse exists.



ADM

PROP

INT

COMP

VER

YOU CANNOT CANCEL THE PAST

REVERSIBLE

Draft created

Delete it. A genuine inverse exists.

COMPENSATABLE

Jira ticket

Close it through deterministic,
pre-approved code, never back
through the agent.

YOU CANNOT CANCEL THE PAST

REVERSIBLE

Draft created

Delete it. A genuine inverse exists.

COMPENSATABLE

Jira ticket

Close it through deterministic, pre-approved code, never back through the agent.

IRREVERSIBLE

Email sent

No true inverse. Ours just sent one.

YOU CANNOT CANCEL THE PAST

REVERSIBLE

Draft created

Delete it. A genuine inverse exists.

COMPENSATABLE

Jira ticket

Close it through deterministic, pre-approved code, never back through the agent.

IRREVERSIBLE

Email sent

No true inverse. Ours just sent one.

Design recovery semantics before the incident, not during it.

FIRE DRILL #2 · SIMULATED · SAME FIFTEEN SECONDS

Containment-aware shutdown

DRILL 1 · EXECUTION 47

DRILL 2 · EXECUTION 62

Simulated: same workload and timings as Fire Drill #1 · mocked dependencies

FIRE DRILL #2 · SIMULATED · SAME FIFTEEN SECONDS

Containment-aware shutdown

DRILL 1 · EXECUTION 47

T+0 KILL SWITCH DISABLED

T+0 no cancel signal

T+1s new execution REJECTED

DRILL 2 · EXECUTION 62

T+0 KILL SWITCH DISABLED

T+0 CANCEL_REQUESTED · gen 18 → 19

T+1s new execution REJECTED

Simulated: same workload and timings as Fire Drill #1 · mocked dependencies

FIRE DRILL #2 · SIMULATED · SAME FIFTEEN SECONDS

Containment-aware shutdown

DRILL 1 · EXECUTION 47

T+0 KILL SWITCH DISABLED

T+0 no cancel signal

T+1s new execution REJECTED

T+15 email.send SENT

DRILL 2 · EXECUTION 62

T+0 KILL SWITCH DISABLED

T+0 CANCEL_REQUESTED · gen 18 → 19

T+1s new execution REJECTED

T+15 18 ≠ 19 STALE GENERATION

Simulated: same workload and timings as Fire Drill #1 · mocked dependencies

FIRE DRILL #2 · SIMULATED · SAME FIFTEEN SECONDS

Containment-aware shutdown

DRILL 1 · EXECUTION 47

T+0 KILL SWITCH DISABLED

T+0 no cancel signal

T+1s new execution REJECTED

T+15 email.send SENT

T+16s n/a

T+19s n/a

T+24s n/a

DRILL 2 · EXECUTION 62

T+0 KILL SWITCH DISABLED

T+0 CANCEL_REQUESTED · gen 18 → 19

T+1s new execution REJECTED

T+15 18 ≠ 19 STALE GENERATION

T+16s reasoning loop CANCELLED

T+19s jira.close CONFIRMED

T+24s quiet(W=5s) SATISFIED

Simulated: same workload and timings as Fire Drill #1 · mocked dependencies

FIRE DRILL #2 · SIMULATED · SAME FIFTEEN SECONDS

Containment-aware shutdown

DRILL 1 · EXECUTION 47

T+0 KILL SWITCH DISABLED

T+0 no cancel signal

T+1s new execution REJECTED

T+15 email.send **SENT**

T+16s n/a

T+19s n/a

T+24s n/a

NOT DEMONSTRATED

DRILL 2 · EXECUTION 62

T+0 KILL SWITCH DISABLED

T+0 **CANCEL_REQUESTED · gen 18 → 19**

T+1s new execution REJECTED

T+15 18 $\neq$ 19 **STALE GENERATION**

T+16s reasoning loop **CANCELLED**

T+19s jira.close **CONFIRMED**

T+24s quiet(W=5s) **SATISFIED**

CONTAINED · TTC 00:24

The queue didn't get faster. The worker stopped trusting stale authority.

Simulated: same workload and timings as Fire Drill #1 · mocked dependencies

Containment you can point at

invoke_agent

chat

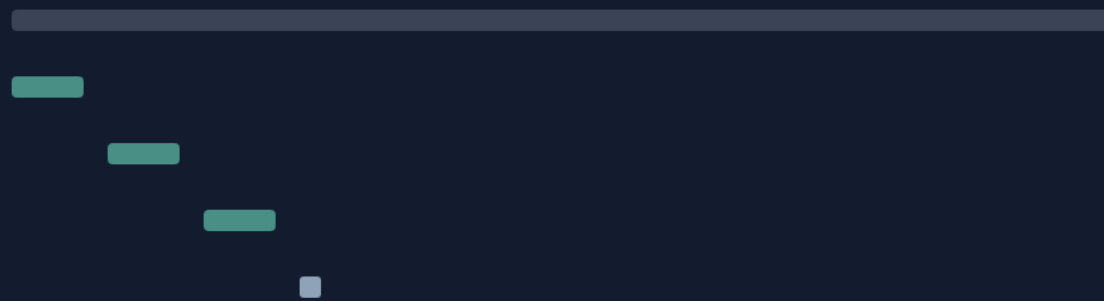
crm.tool

jira.tool

enqueue

worker

compensate



PROVE IT · EXECUTION 62 · ONE TRACE

Containment you can point at

invoke_agent

chat

crm.tool

jira.tool

enqueue

worker

compensate

CANCEL REQUESTED

T+0 · compensation starts

Containment you can point at

invoke_agent

chat

crm.tool

jira.tool

enqueue

worker

compensate

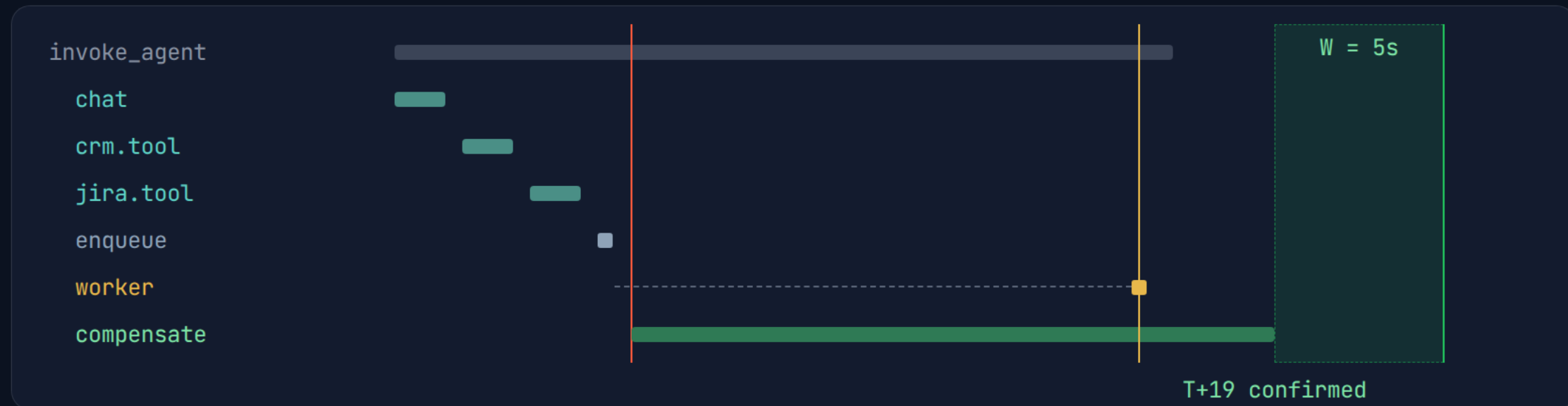
CANCEL REQUESTED

T+0 · compensation starts

STALE GENERATION

T+15 · 18 $\neq$ 19 · refused

Containment you can point at



CANCEL REQUESTED

`T+0 · compensation starts`

STALE GENERATION

`T+15 · 18  $\neq$  19 · refused`

CONTAINED

`T+24 · W = 5s satisfied`

TIME TO CONTAINMENT

THIS TALK'S METRIC · NOT AN INDUSTRY STANDARD

Disabled is not an SLI

CONTAINED =

`0 active + 0 unresolved + quiet(W)`

TTC = `t(contained)` - `t(kill)`

`W = 5 seconds in this drill · workload-specific`

TIME TO CONTAINMENT

THIS TALK'S METRIC · NOT AN INDUSTRY STANDARD

Disabled is not an SLI

CONTAINED =

0 active + 0 unresolved + quiet(W)

TTC = t(contained) - t(kill)

W = 5 seconds in this drill · workload-specific

FIRE DRILL 1 · TIME TO CONTAINMENT

NOT DEMONSTRATED

1 unauthorised side effect, irreversible

0 compensations · email already sent

TIME TO CONTAINMENT

THIS TALK'S METRIC · NOT AN INDUSTRY STANDARD

Disabled is not an SLI

CONTAINED =

0 active + 0 unresolved + quiet(W)

$TTC = t(\text{contained}) - t(\text{kill})$

W = 5 seconds in this drill · workload-specific

FIRE DRILL 1 · TIME TO CONTAINMENT

NOT DEMONSTRATED

1 unauthorised side effect, irreversible

0 compensations · email already sent

FIRE DRILL 2 · TIME TO CONTAINMENT

24s

0 unauthorised side effects

1 compensation, confirmed · 0 unresolved

RED-TEAM THE SHUTDOWN PATH

ADMISSION stale flag state · SDK can't reach provider, fails open

PROPAGATION disconnected worker, reconciler also down

RED-TEAM THE SHUTDOWN PATH

ADMISSION	stale flag state · SDK can't reach provider, fails open
PROPAGATION	disconnected worker, reconciler also down
INTERRUPTION	call already in flight

RED-TEAM THE SHUTDOWN PATH

ADMISSION	stale flag state · SDK can't reach provider, fails open
PROPAGATION	disconnected worker, reconciler also down
INTERRUPTION	call already in flight
COMPENSATION	the corrective action itself fails

RED-TEAM THE SHUTDOWN PATH

ADMISSION	stale flag state · SDK can't reach provider, fails open
PROPAGATION	disconnected worker, reconciler also down
INTERRUPTION	call already in flight
COMPENSATION	the corrective action itself fails
VERIFICATION	activity outside the boundary you can see

RED-TEAM THE SHUTDOWN PATH

ADMISSION	stale flag state · SDK can't reach provider, fails open
PROPAGATION	disconnected worker, reconciler also down
INTERRUPTION	call already in flight
COMPENSATION	the corrective action itself fails
VERIFICATION	activity outside the boundary you can see

ONE FAILURE ONLY AGENTS ADD

CANCELLATION IS NOT THE MODEL'S CALL

enforce it in the harness, before the model's next turn

RESUME

REPLAY

FLUSH

THREE QUESTIONS

- 1 What happens to work already in flight when the flag changes?

THREE QUESTIONS

1 What happens to work already in flight when the flag changes?

2 Which side effects are interruptible, compensatable, or irreversible?

THREE QUESTIONS

- 1 What happens to work already in flight when the flag changes?
- 2 Which side effects are interruptible, compensatable, or irreversible?
- 3 What evidence proves autonomous activity has actually stopped, within what you can observe?

THREE QUESTIONS

- 1 What happens to work already in flight when the flag changes?
- 2 Which side effects are interruptible, compensatable, or irreversible?
- 3 What evidence proves autonomous activity has actually stopped, within what you can observe?

If you can't answer these three, you have a very good feature flag.

THREE QUESTIONS

- 1 What happens to work already in flight when the flag changes?
- 2 Which side effects are interruptible, compensatable, or irreversible?
- 3 What evidence proves autonomous activity has actually stopped, within what you can observe?

If you can't answer these three, you have a very good feature flag.

You don't have a kill switch.

Disabled is a configuration state.

Ask what happens fifteen seconds later.

Disabled is a configuration state.

Ask what happens fifteen seconds later.

Contained is a runtime property.