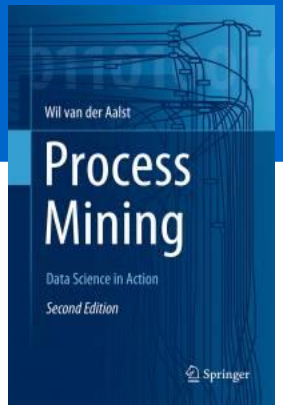
Process Mining: Data Science in Action

Conformance Checking Using Token-Based Replay

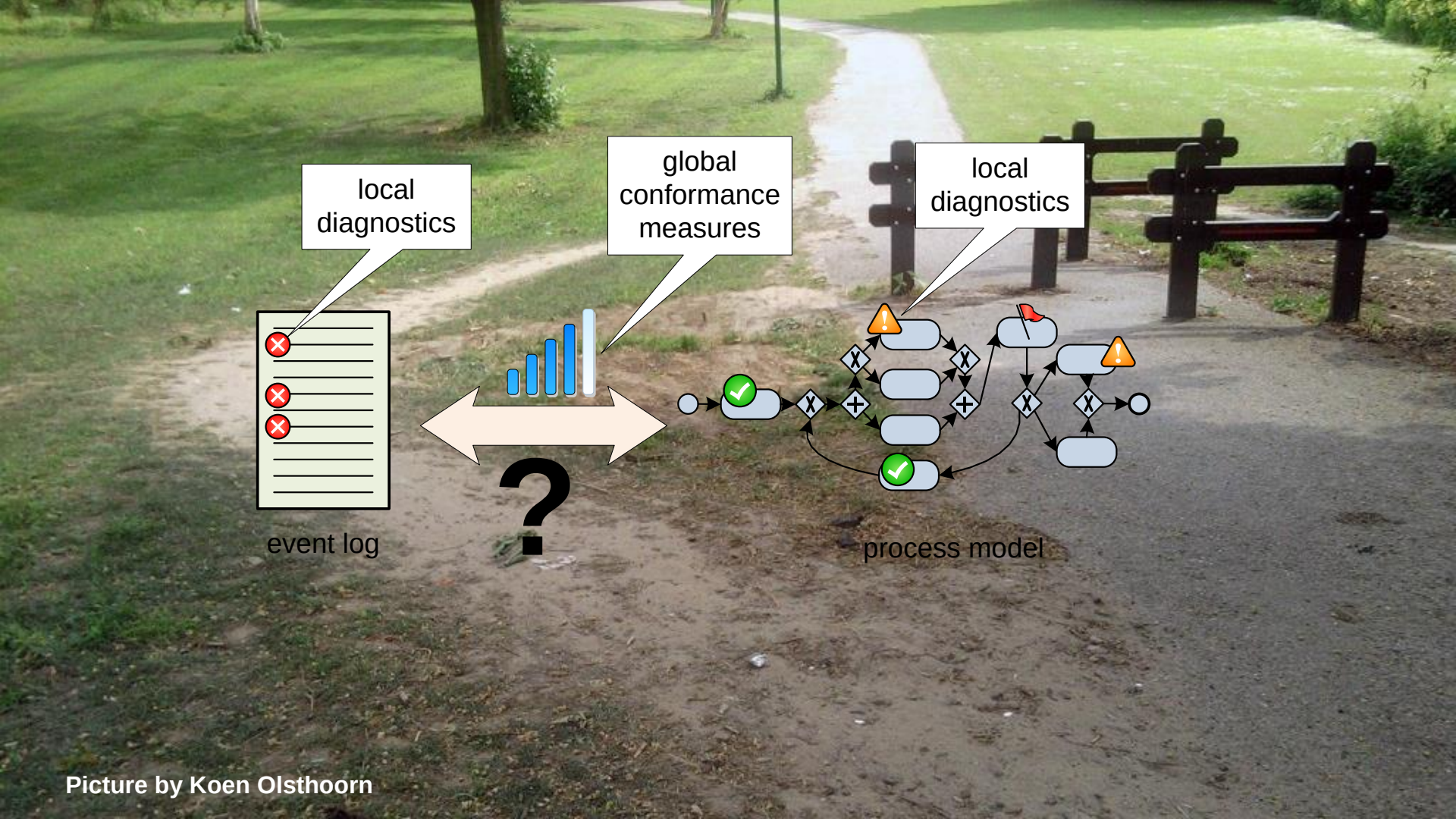


prof.dr.ir. Wil van der Aalst
www.processmining.org

TU/e Technische Universiteit Eindhoven University of Technology

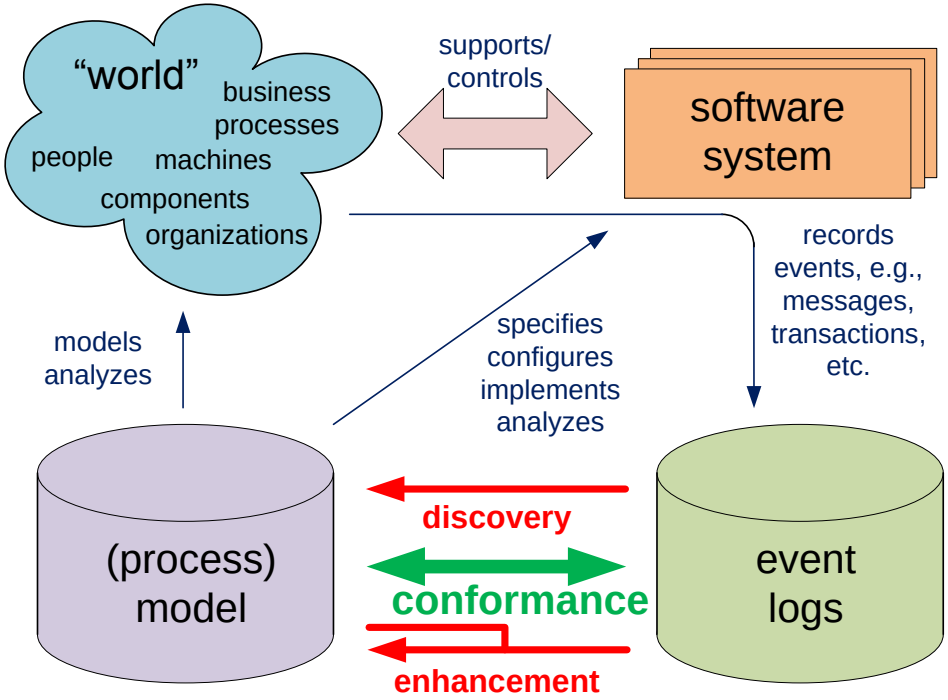
Where innovation starts

Picture by Koen Olsthoorn



Picture by Koen Olsthoorn

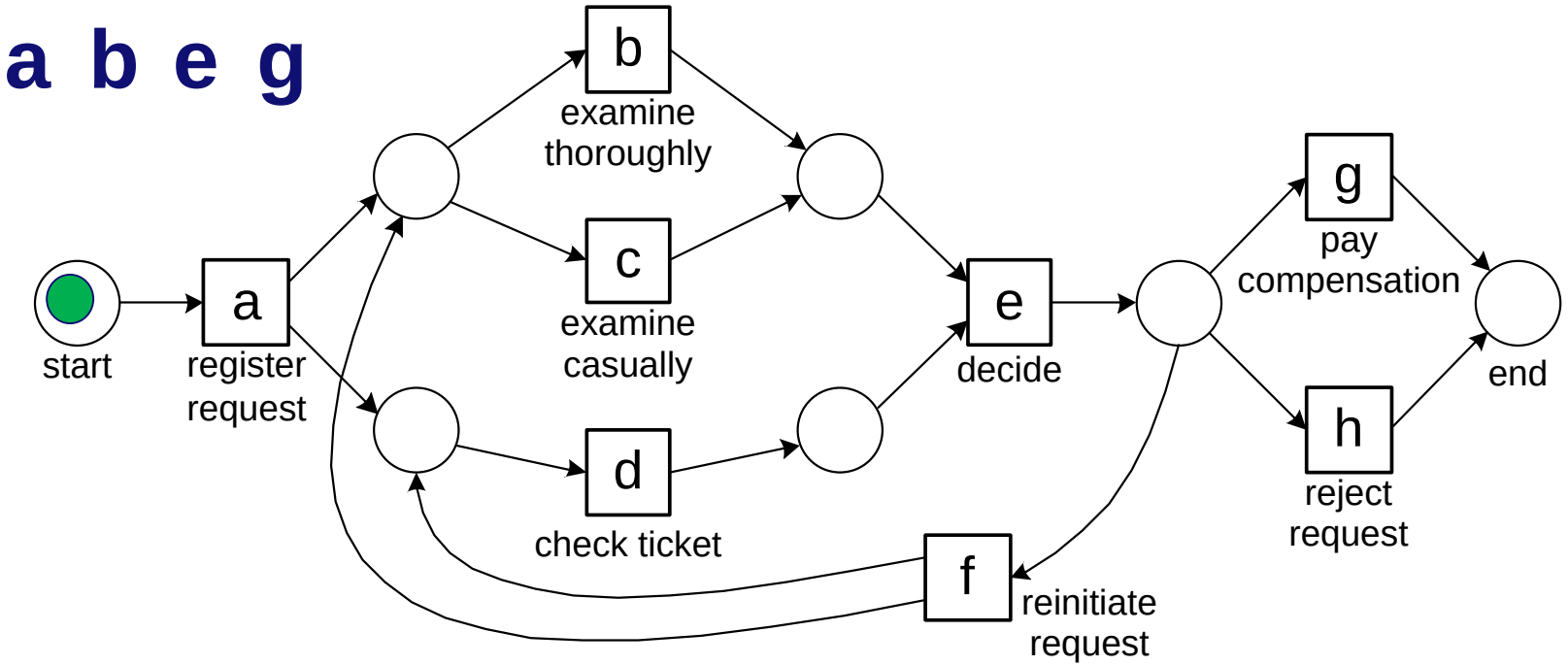
Conformance checking



1. Conformance checking using causal footprints.
2. **Conformance checking based on token-based replay.**
3. Alignment-based conformance checking.

TU/e

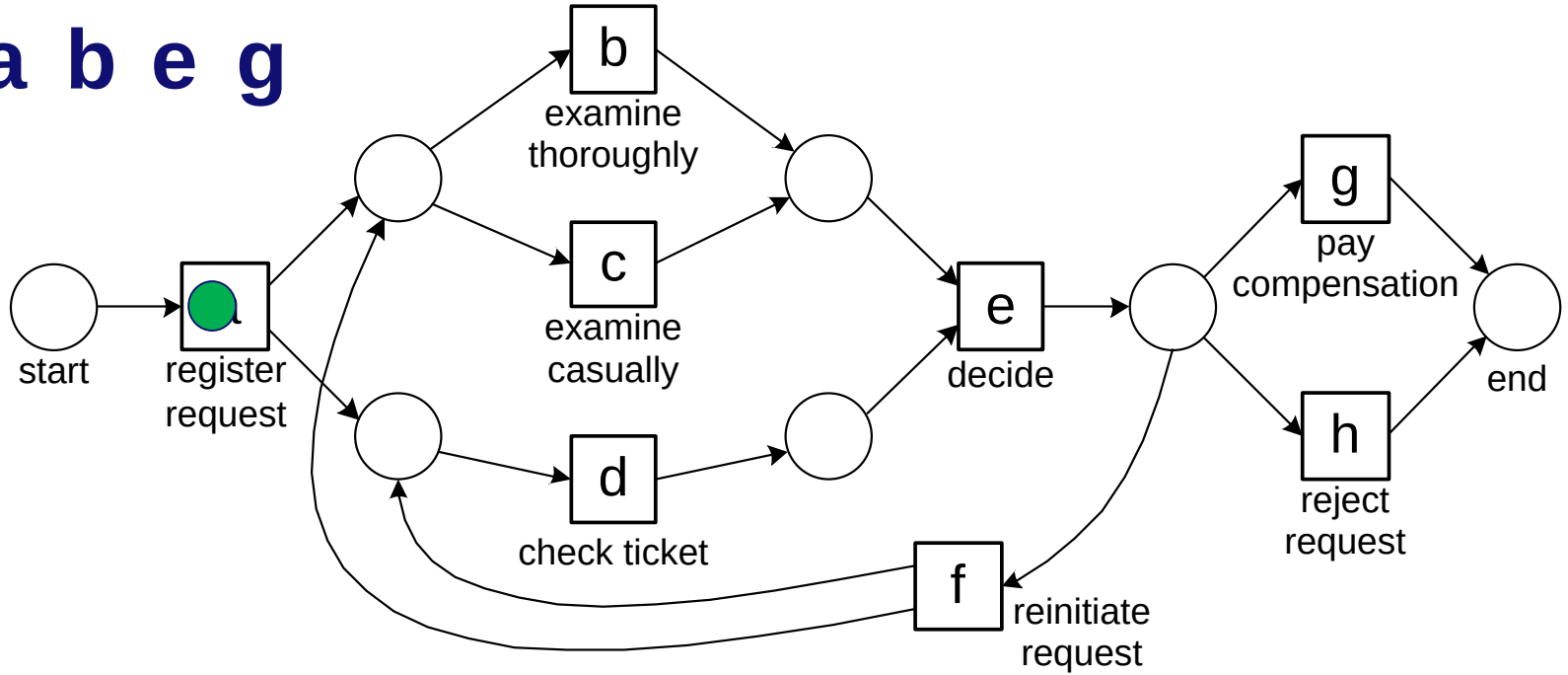
Counting tokens while replaying



TU/e

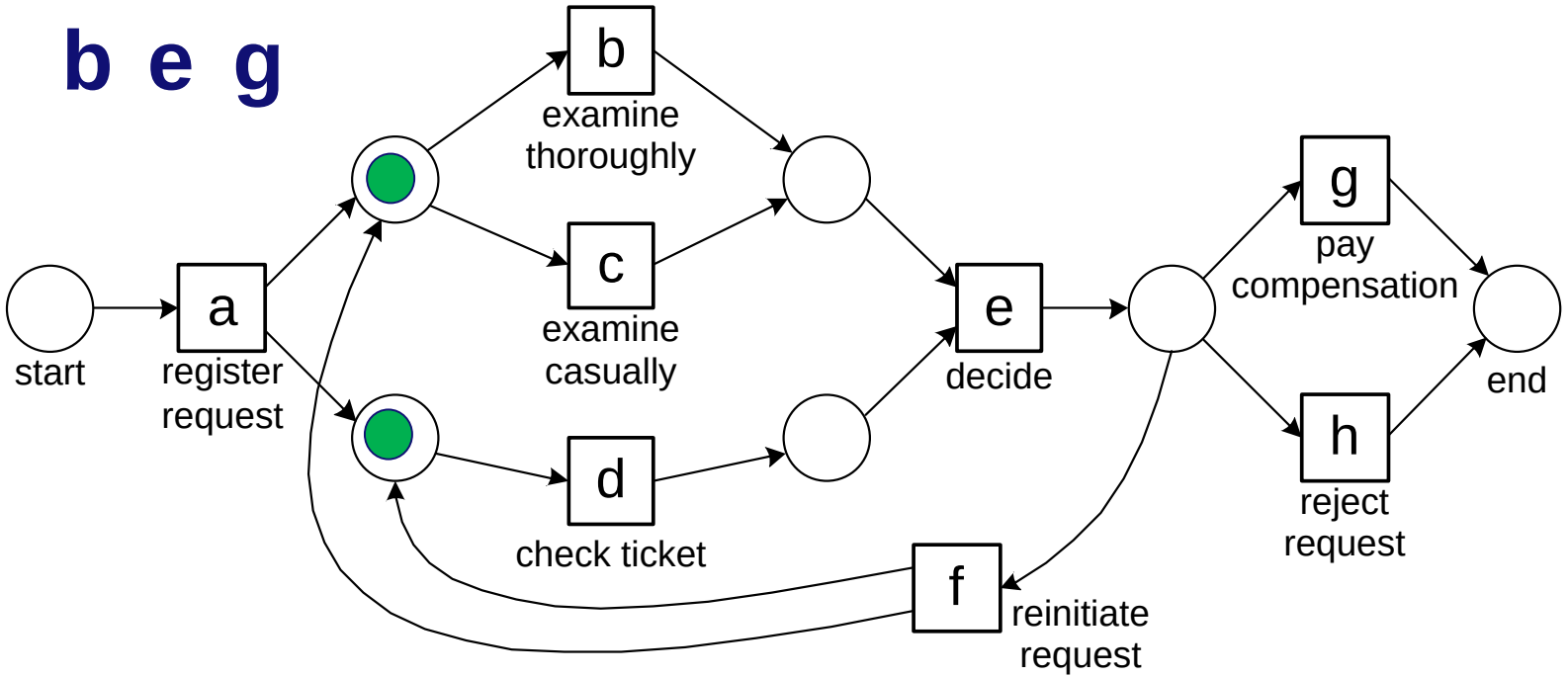
Counting tokens while replaying

a b e g



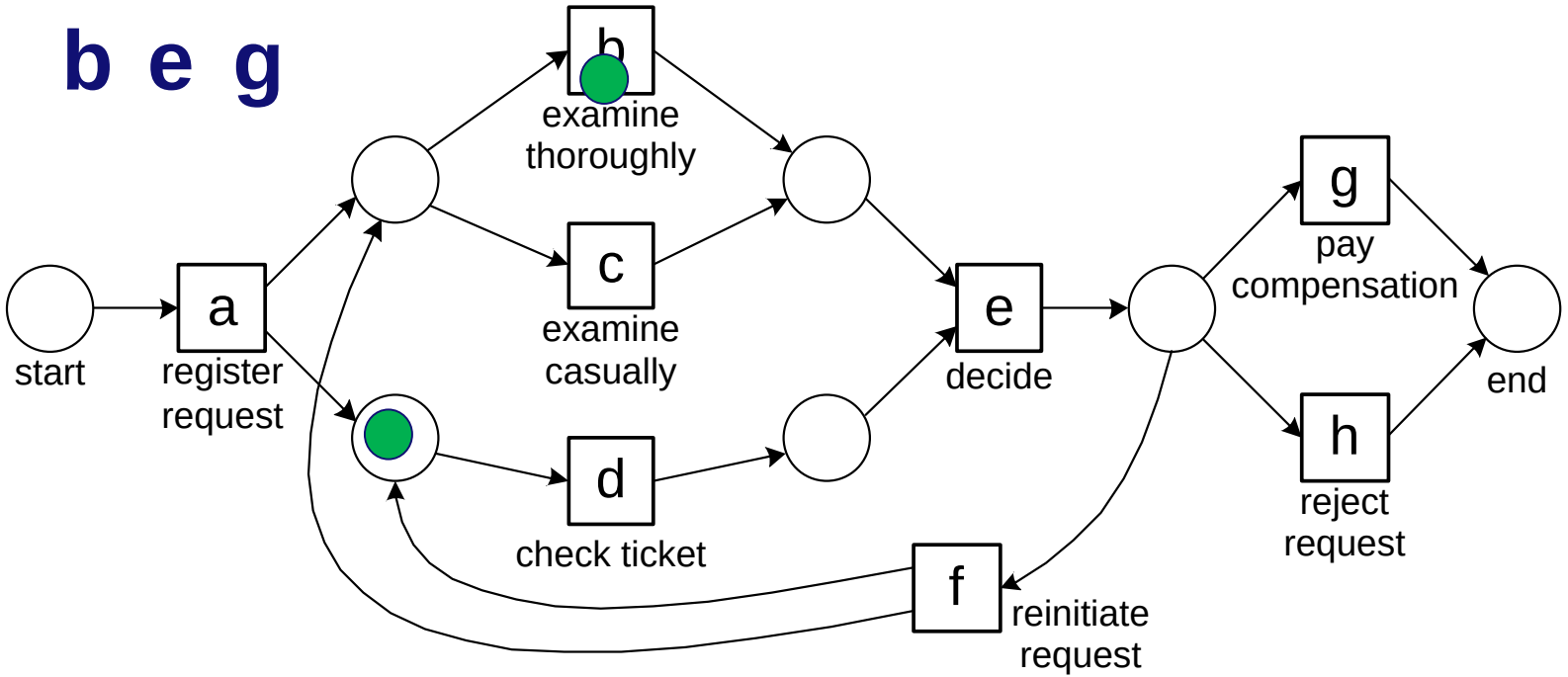
TU/e

Counting tokens while replaying

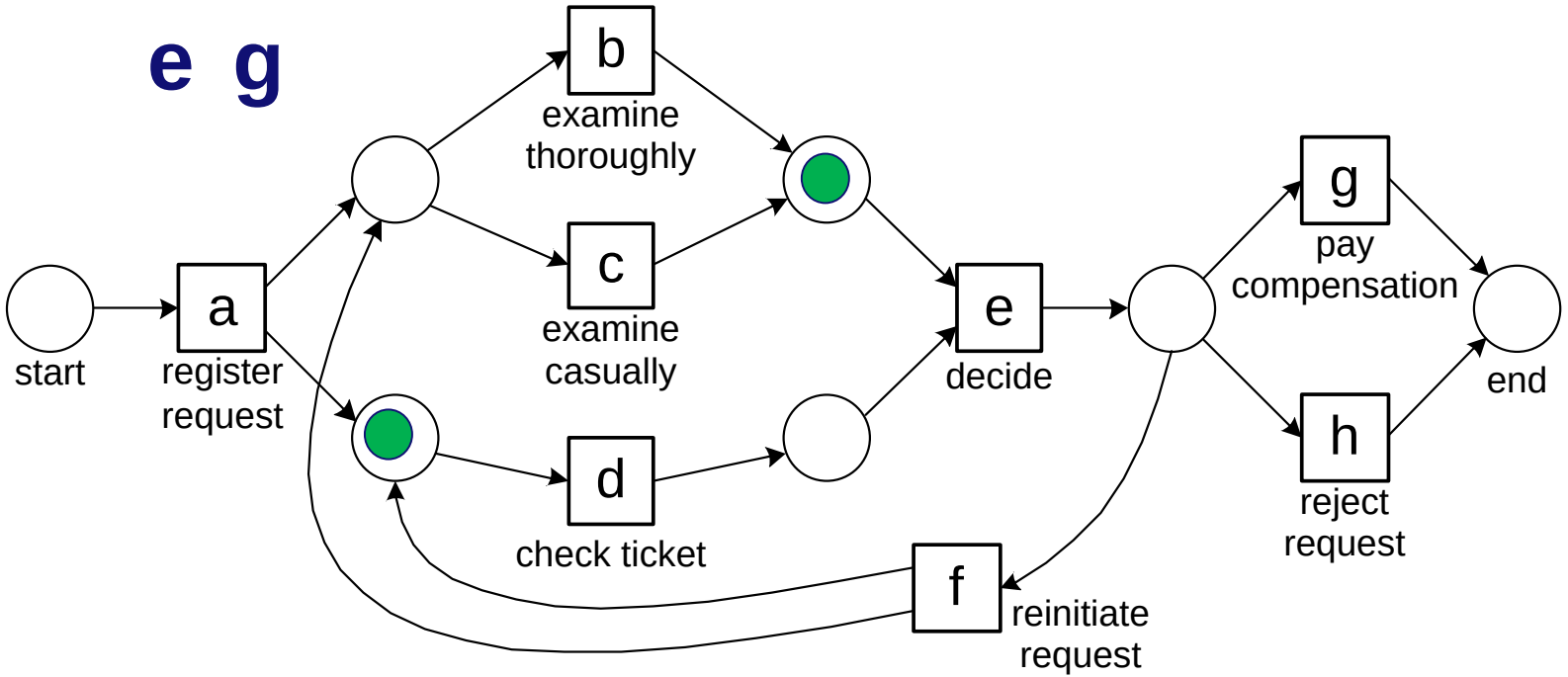


TU/e

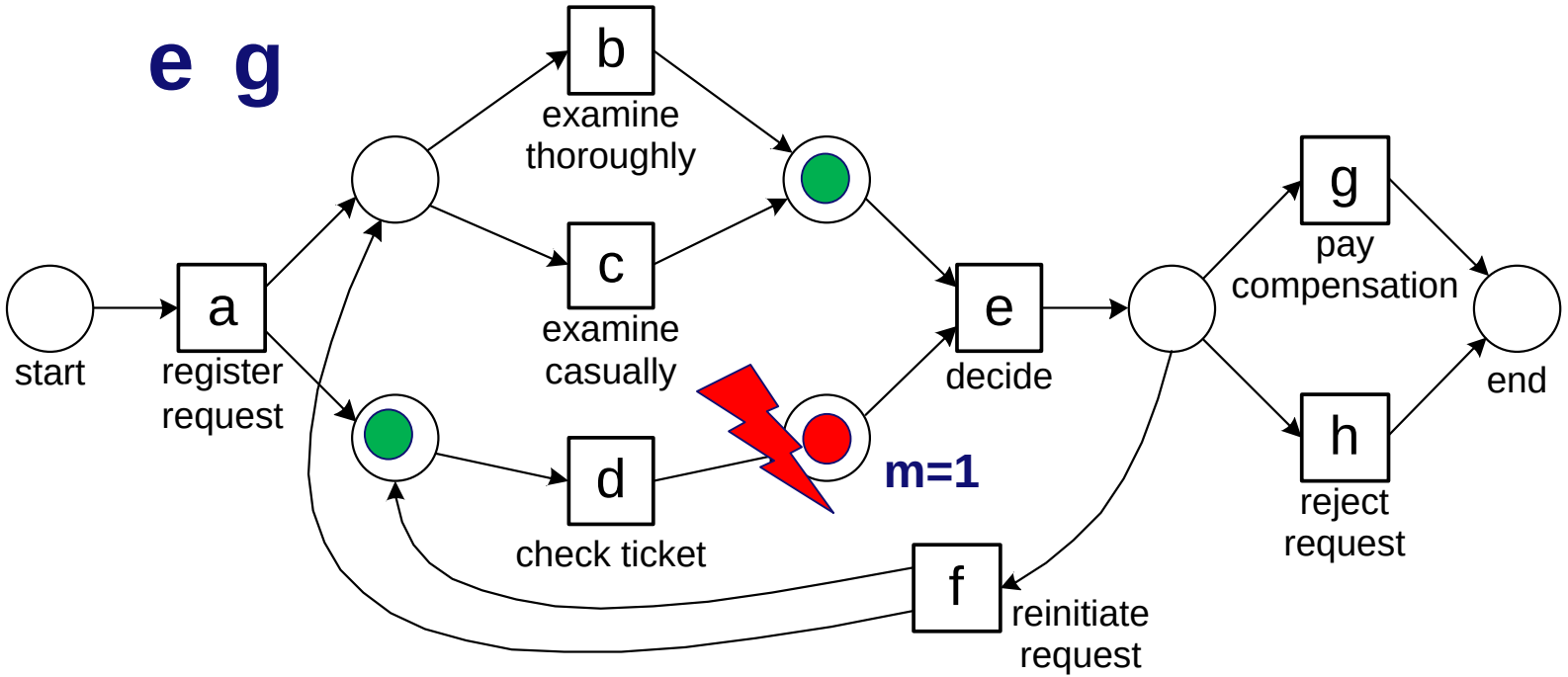
Counting tokens while replaying



Counting tokens while replaying

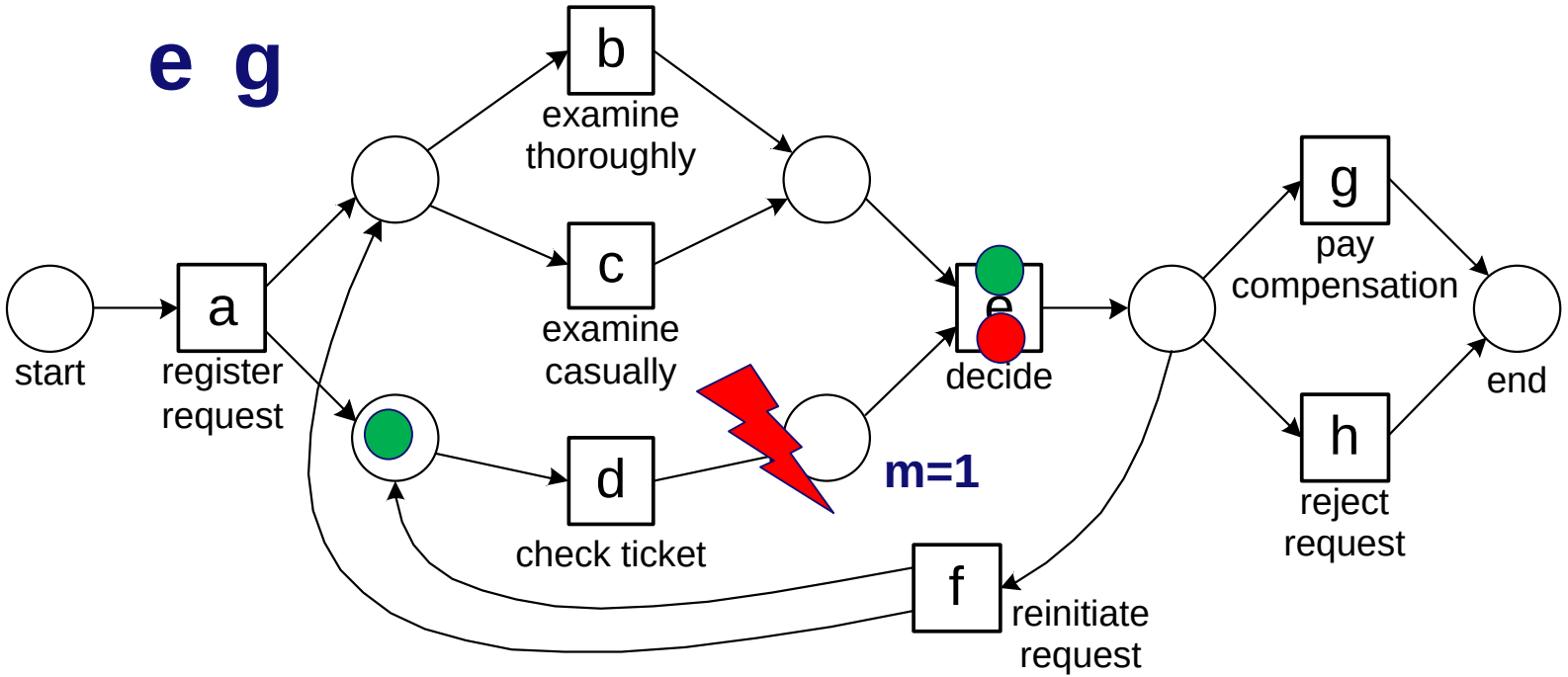


Counting tokens while replaying

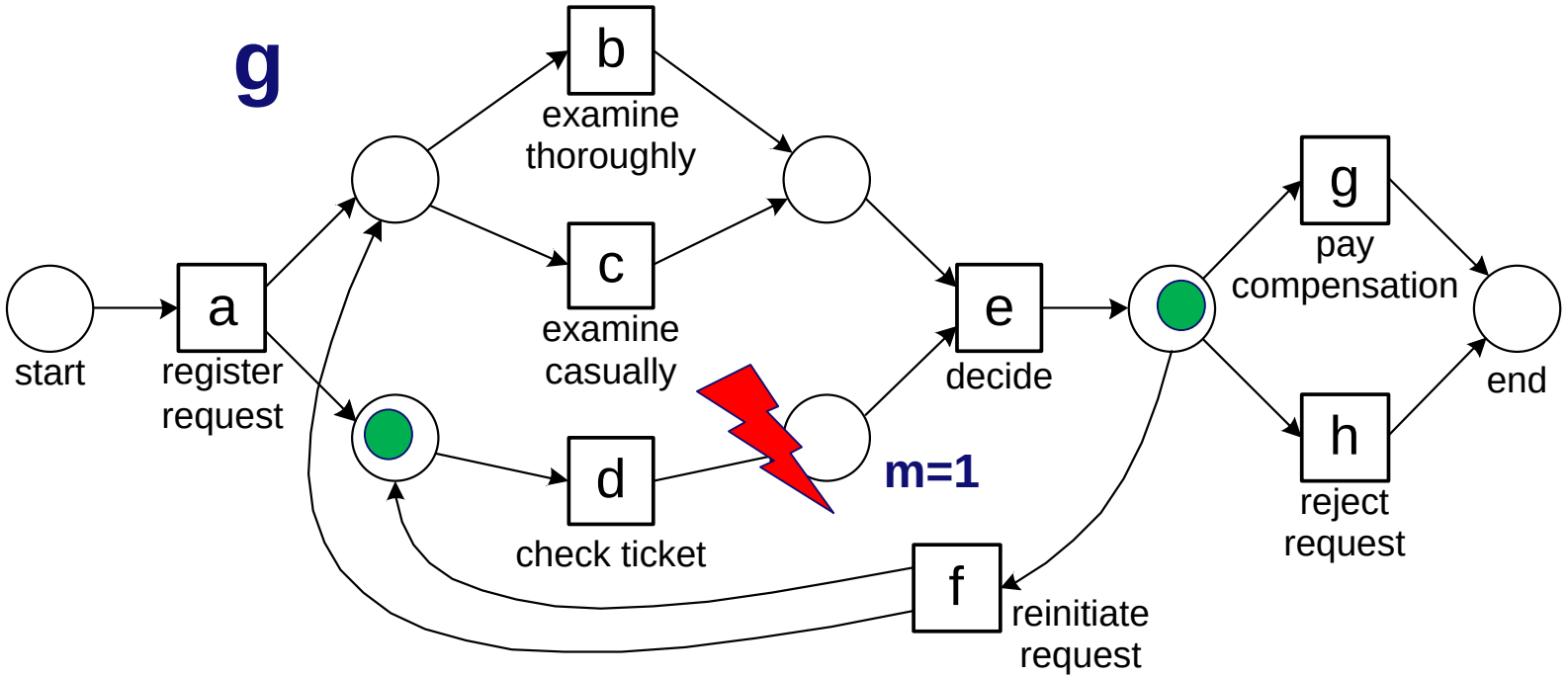


TU/e

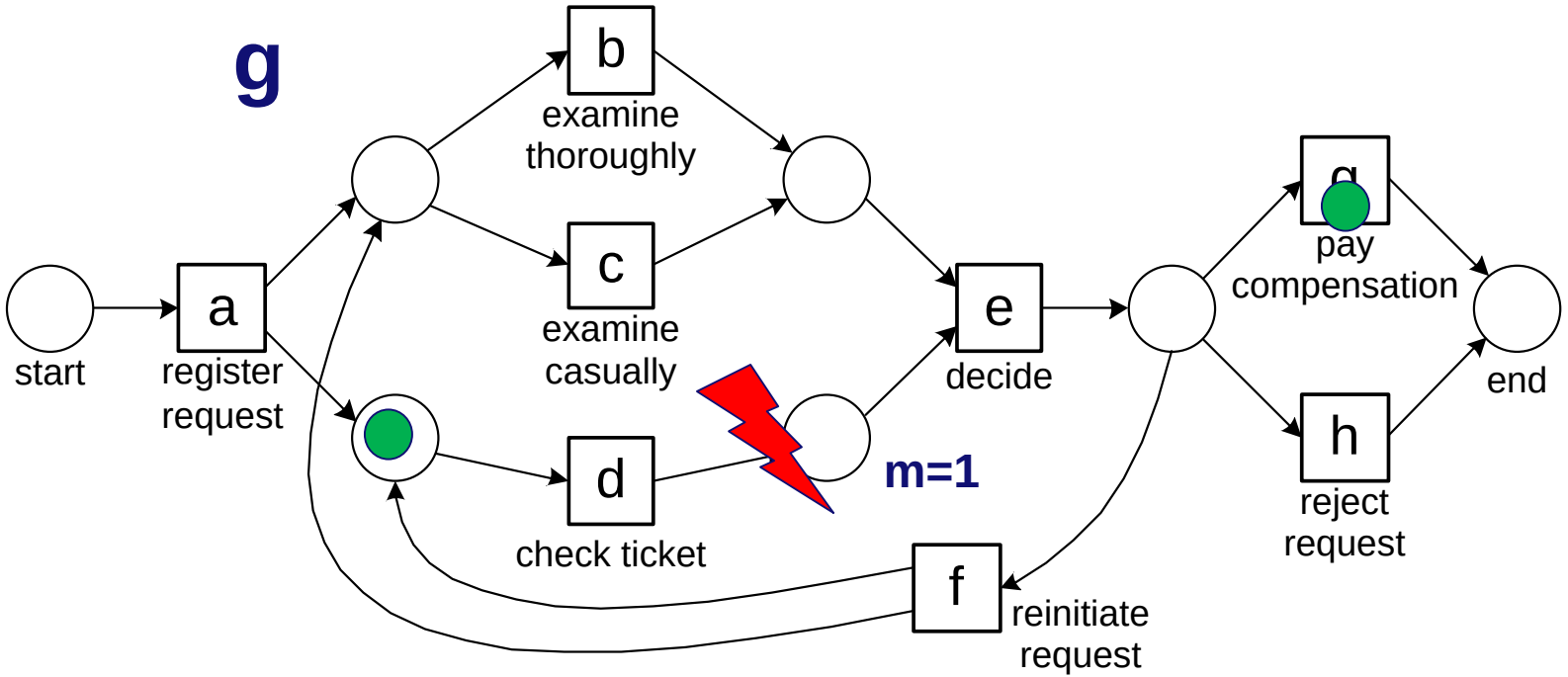
Counting tokens while replaying



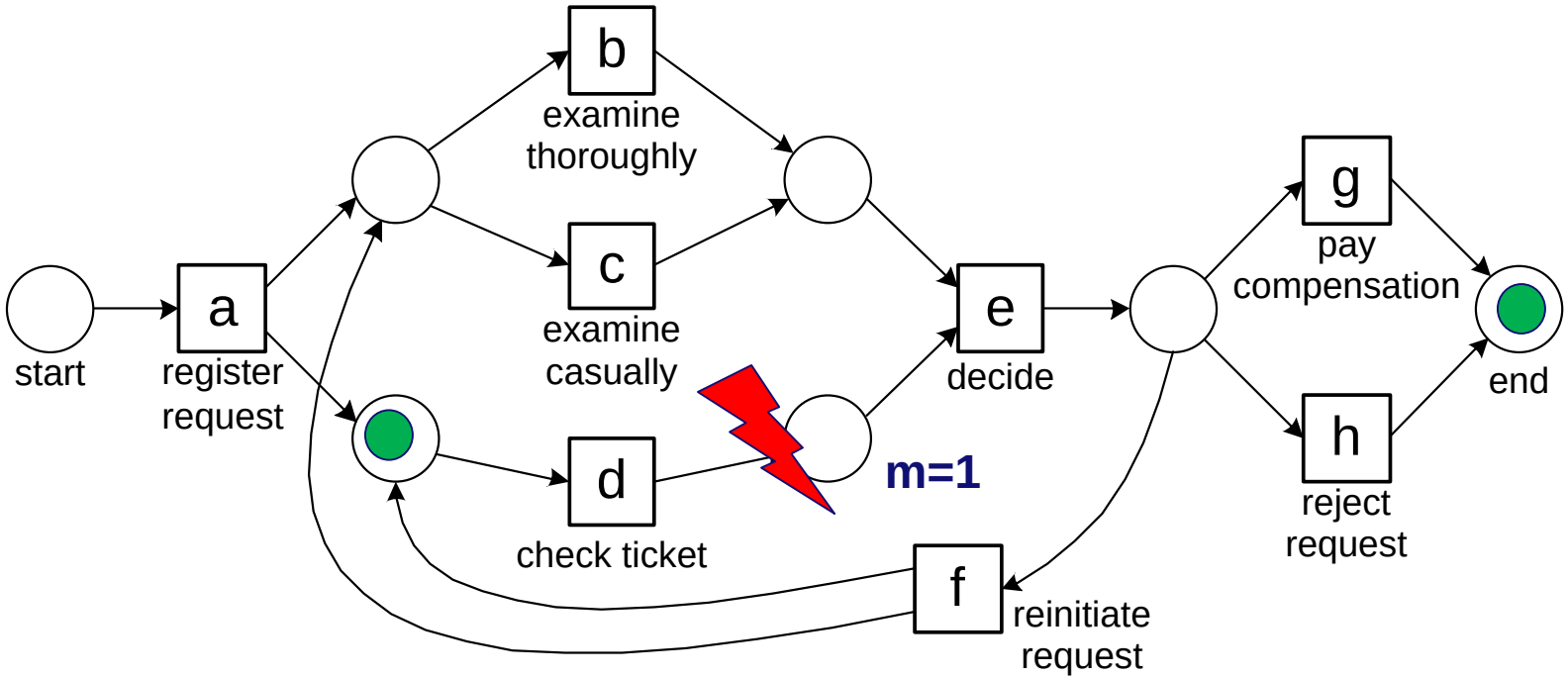
Counting tokens while replaying



Counting tokens while replaying

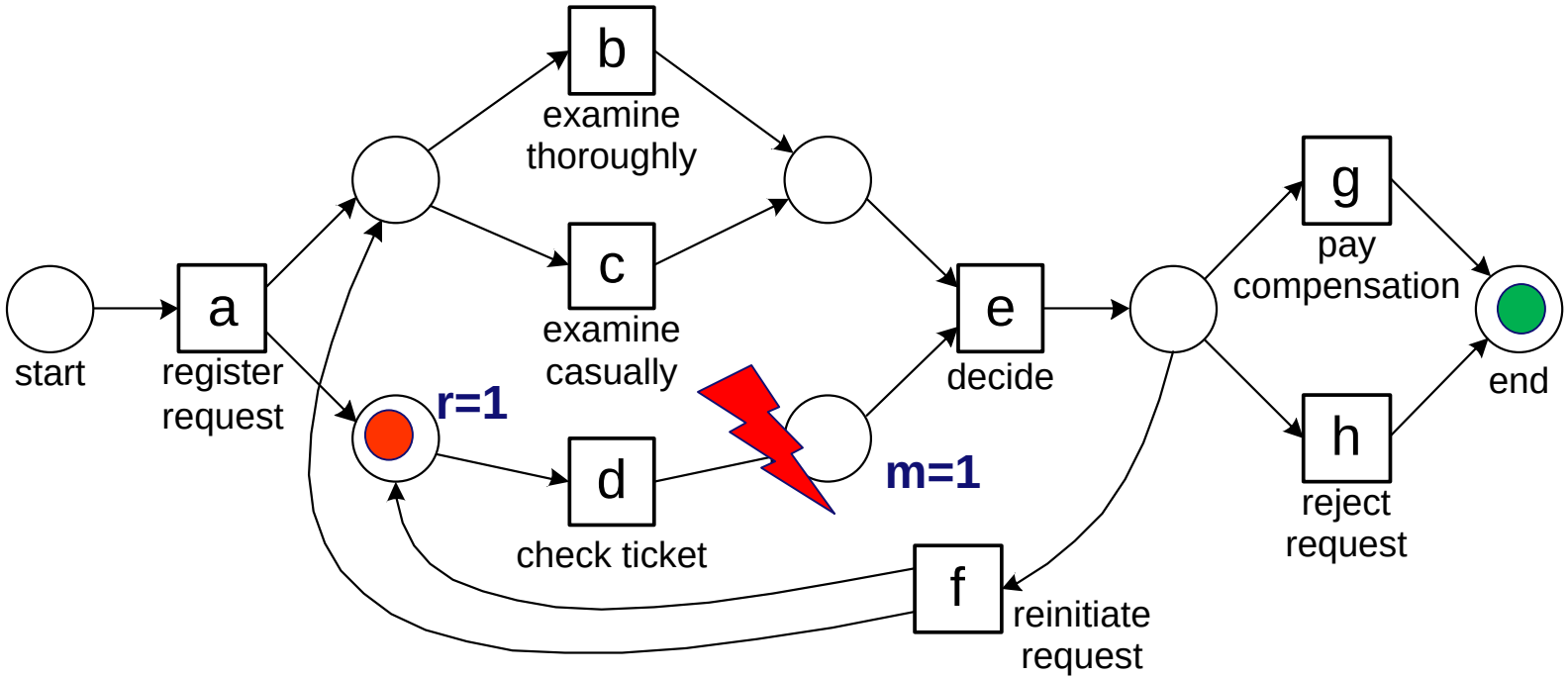


Counting tokens while replaying

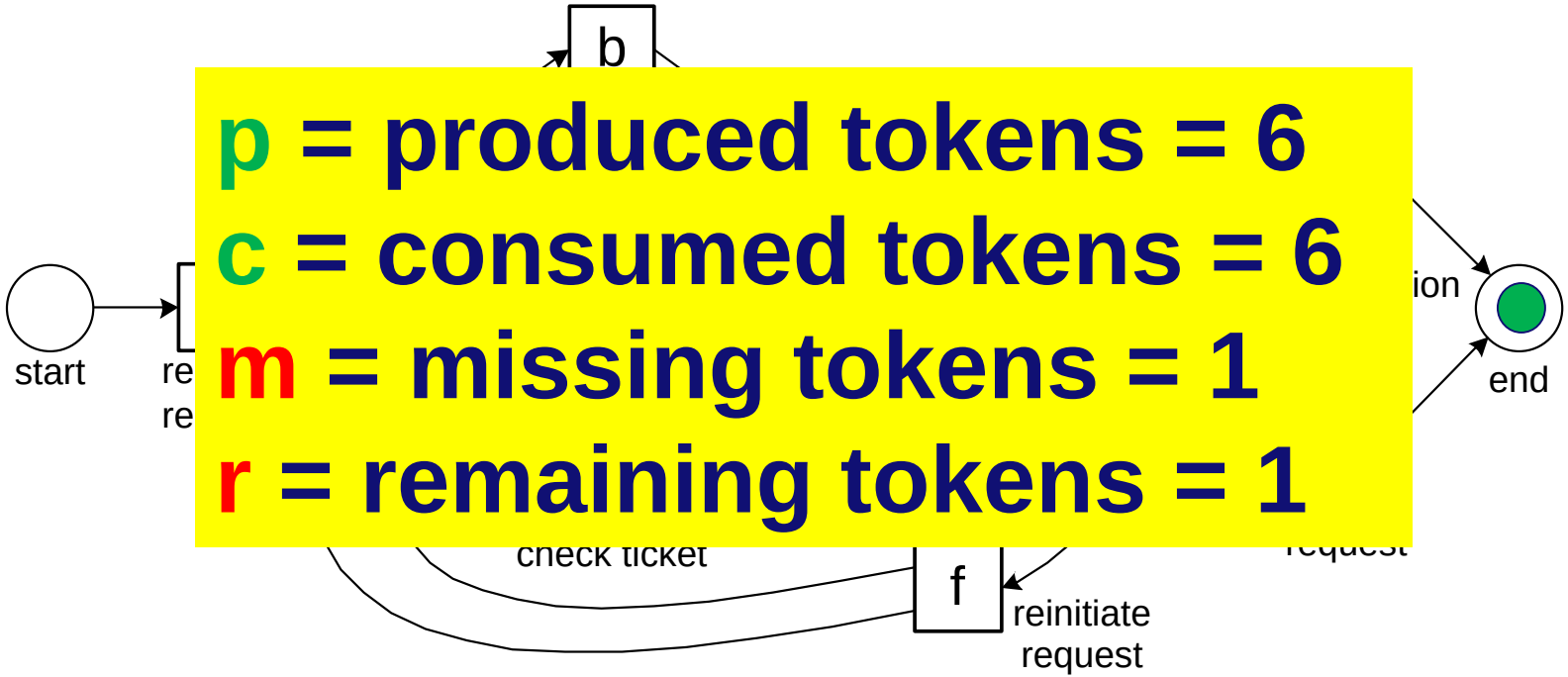


TU/e

Counting tokens while replaying



Counting tokens while replaying



Quantifying fitness at the trace level

$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

p = produced tokens = 6
c = consumed tokens = 6
m = missing tokens = 1
r = remaining tokens = 1

TU/e

Quantifying fitness at the trace level

$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right)$$

p = produced tokens = 6
c = consumed tokens = 6
m = missing tokens = 1
r = remaining tokens = 1

TU/e

Quantifying fitness at the trace level

$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.83333$$

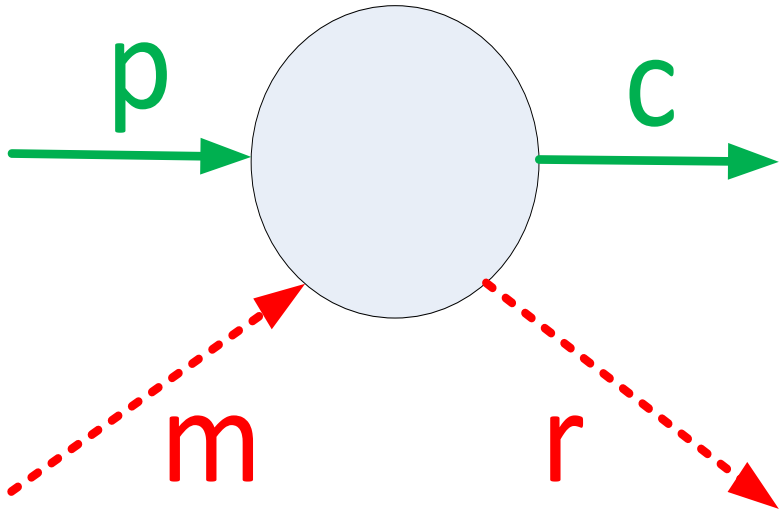
p = produced tokens = 6
c = consumed tokens = 6
m = missing tokens = 1
r = remaining tokens = 1

TU/e

Approach (1/3)

Use four counters:

- **p = produced tokens**
- **c = consumed tokens**
- **m = missing tokens**
  (consumed while not there)
- **r = remaining tokens**
  (produced but not consumed)

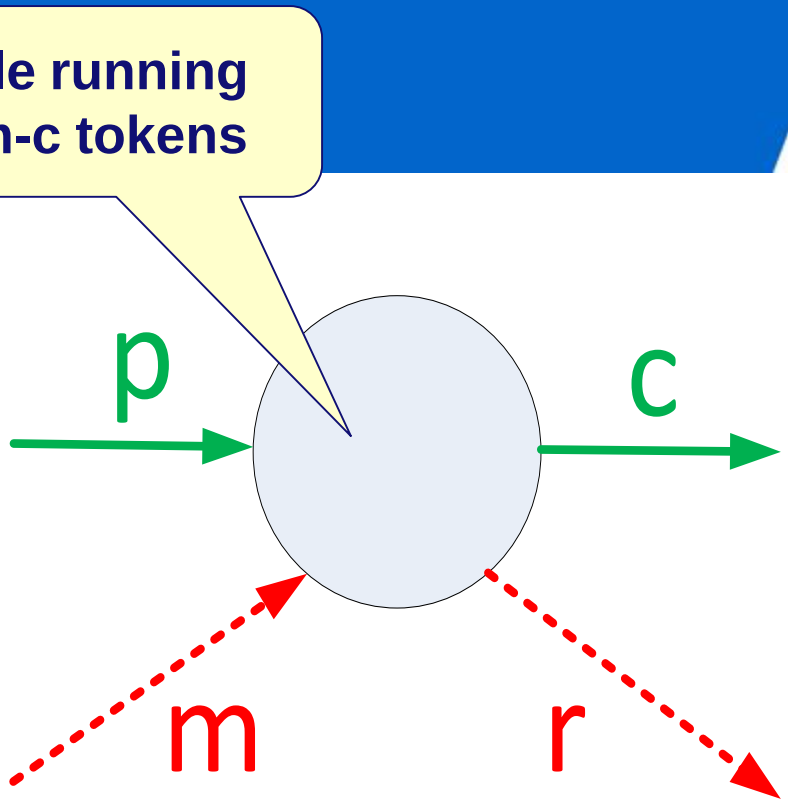


TU/e

Approach (1/3)

Use four counters:

- **p = produced tokens**
- **c = consumed tokens**
- **m = missing tokens**
  (consumed while not there)
- **r = remaining tokens**
  (produced but not consumed)

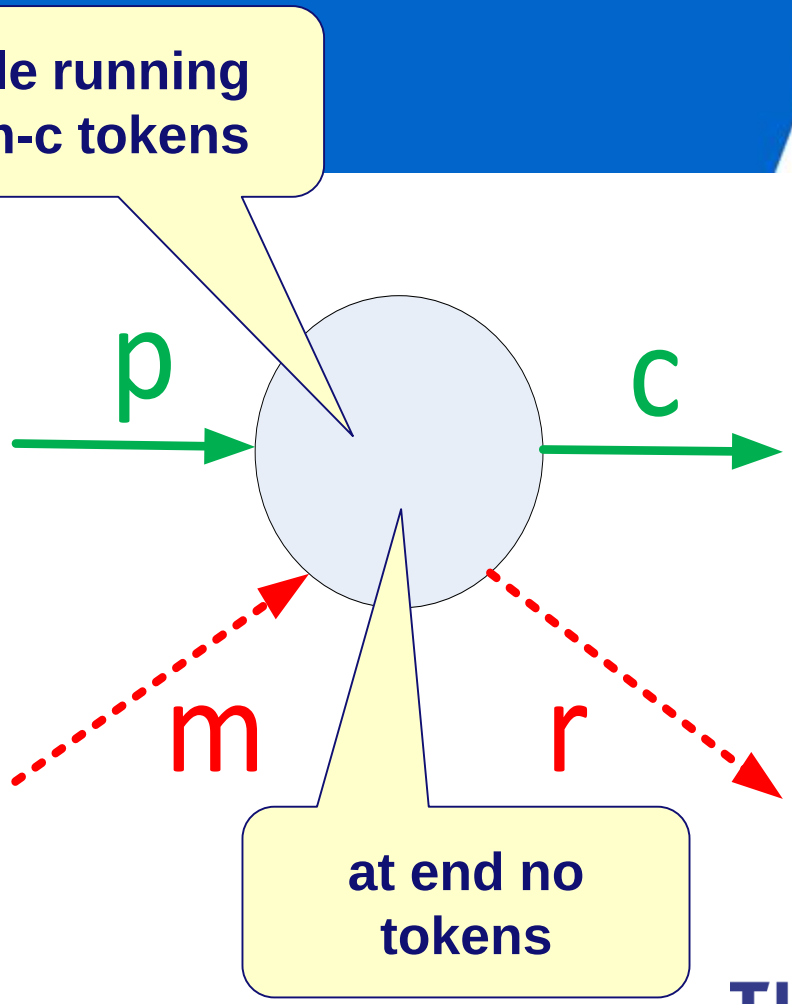


TU/e

Approach (1/3)

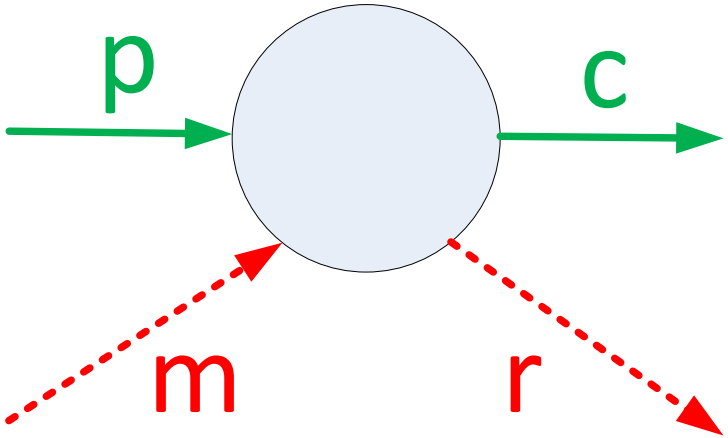
Use four counters:

- p = produced tokens
- c = consumed tokens
- m = missing tokens
  (consumed while not there)
- r = remaining tokens
  (produced but not consumed)



TU/e

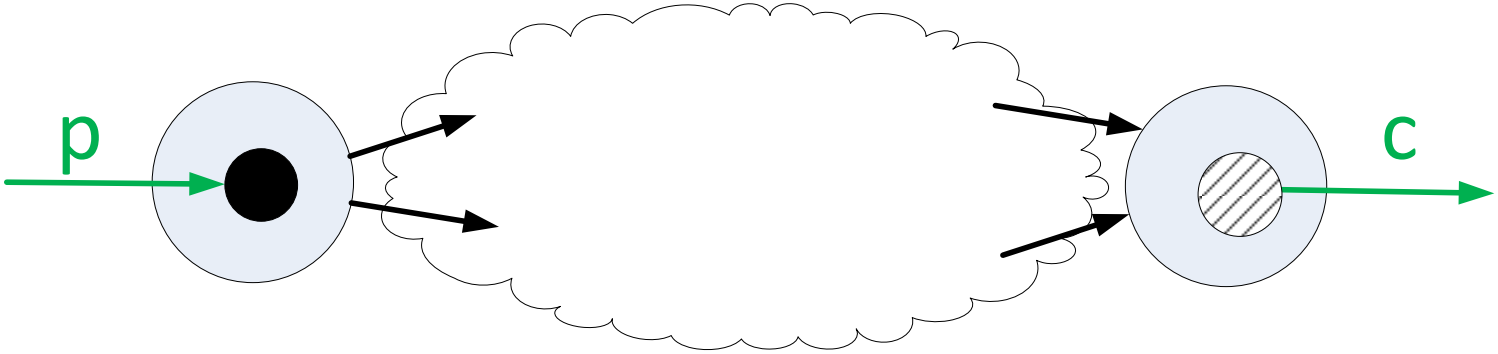
Approach (2/3)



- **Invariants**
  - **At any time: p+m ≥ c ≥ m (also per place)**
  - **At the end: r = p + m - c (also per place)**

TU/e

Approach (3/3)

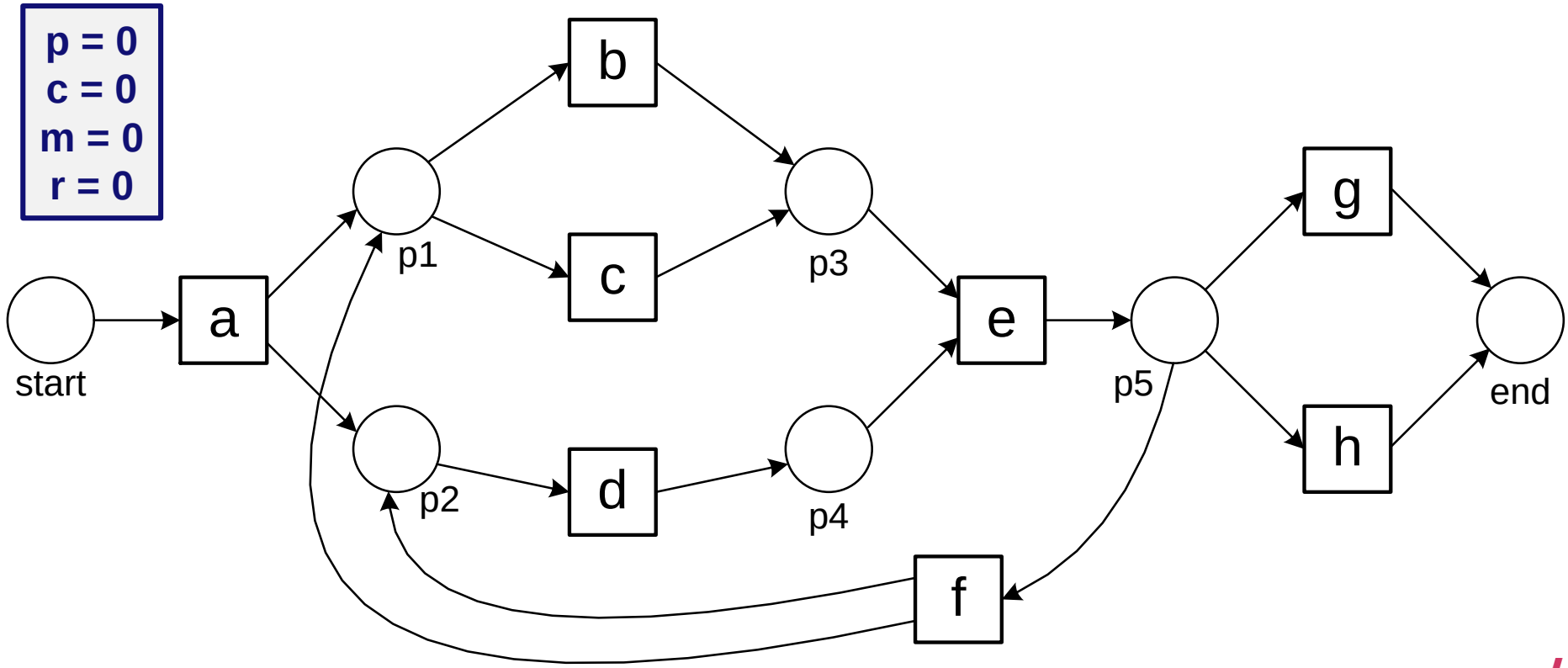


Initialization and finalization:

- **In the beginning a token is produced for the source place: p = 1.**
- **At the end a token is consumed from the sink place (also if not there): c' = c + 1.**

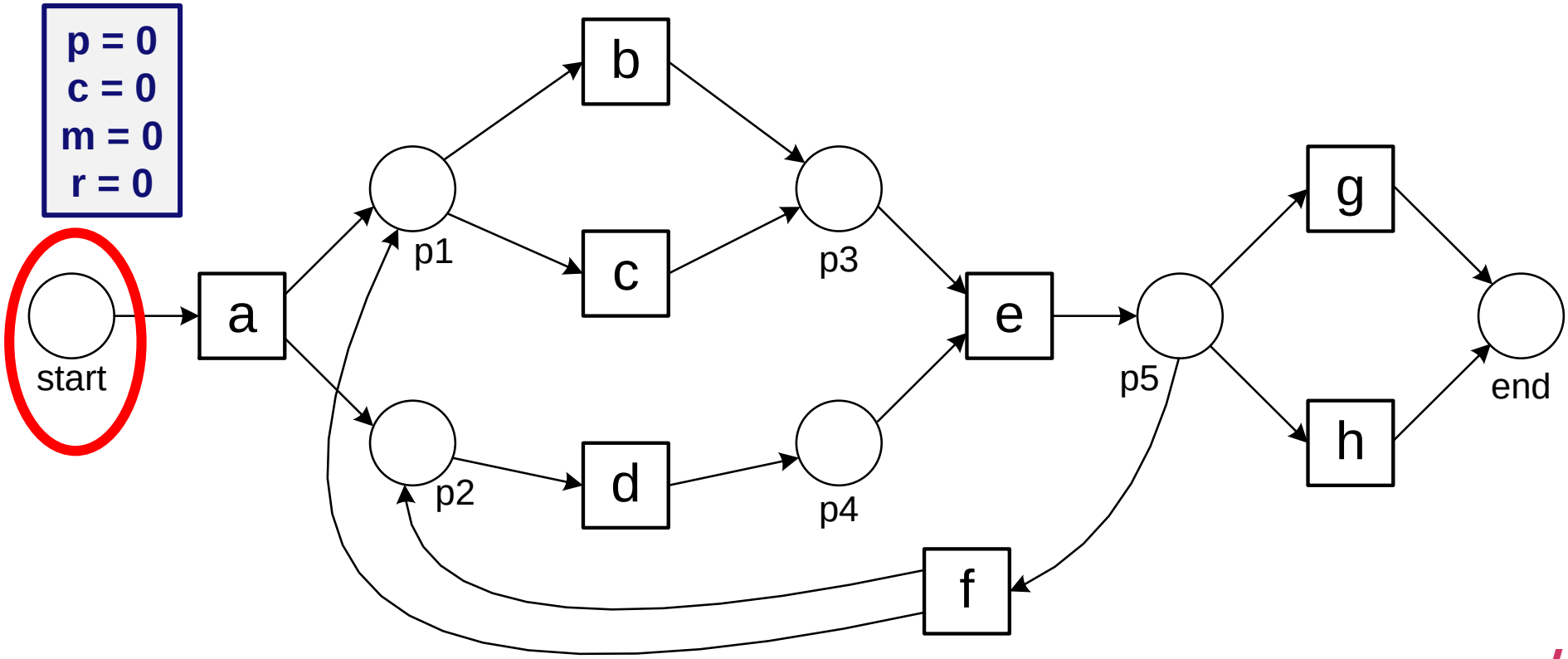
TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



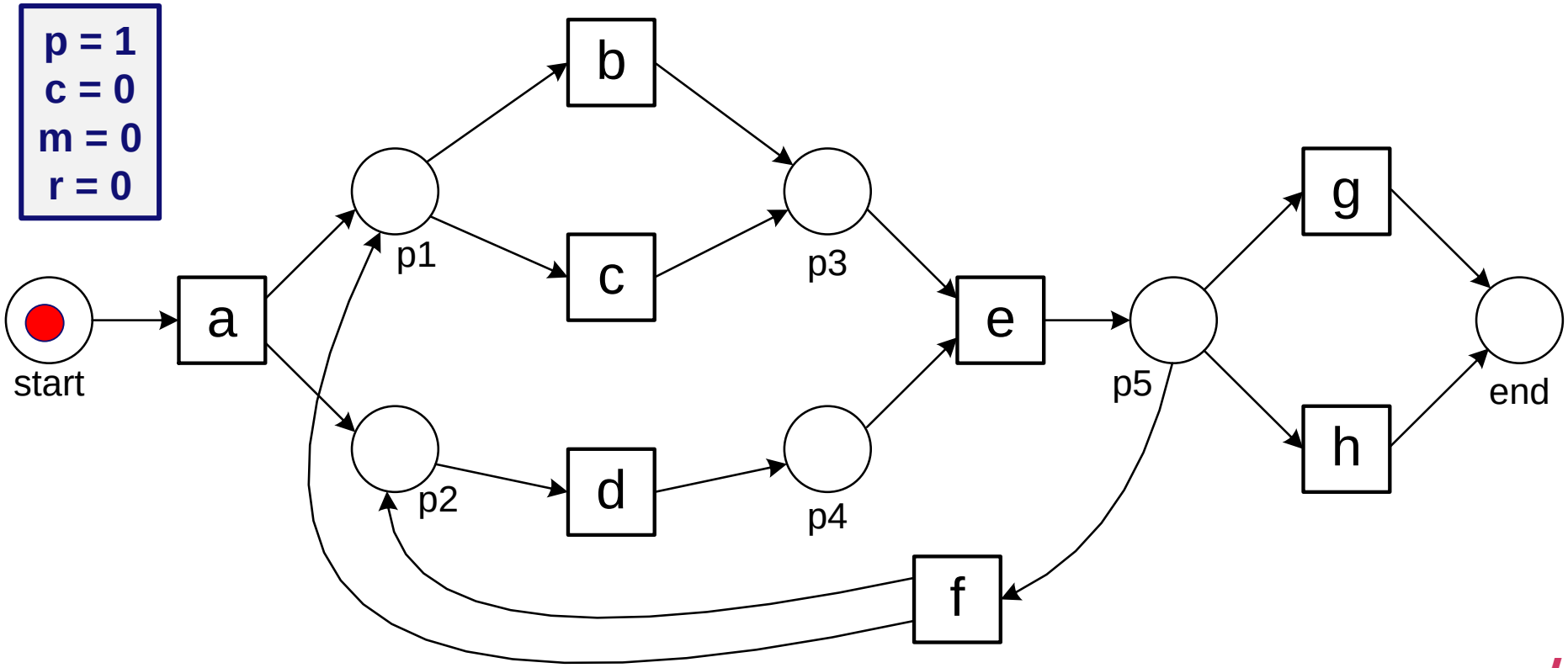
TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



TU/e

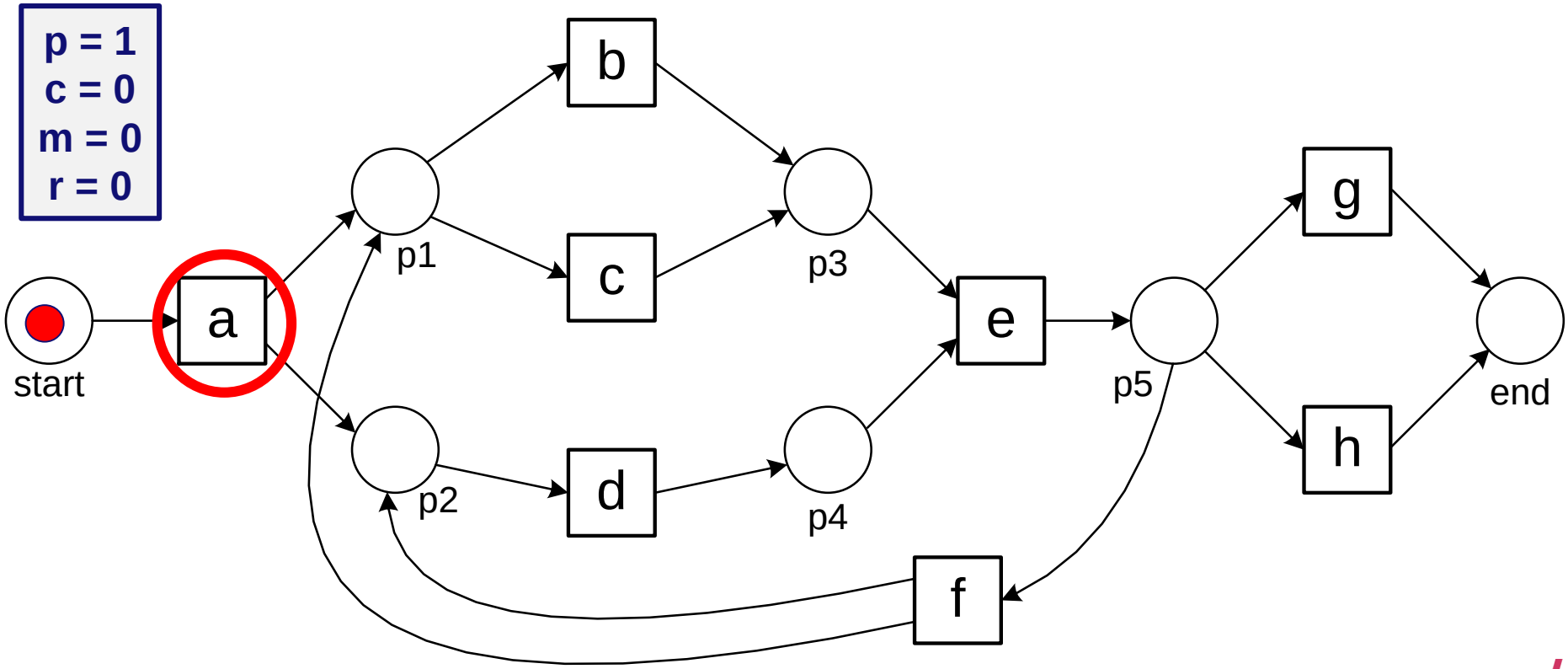
Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



TU/e

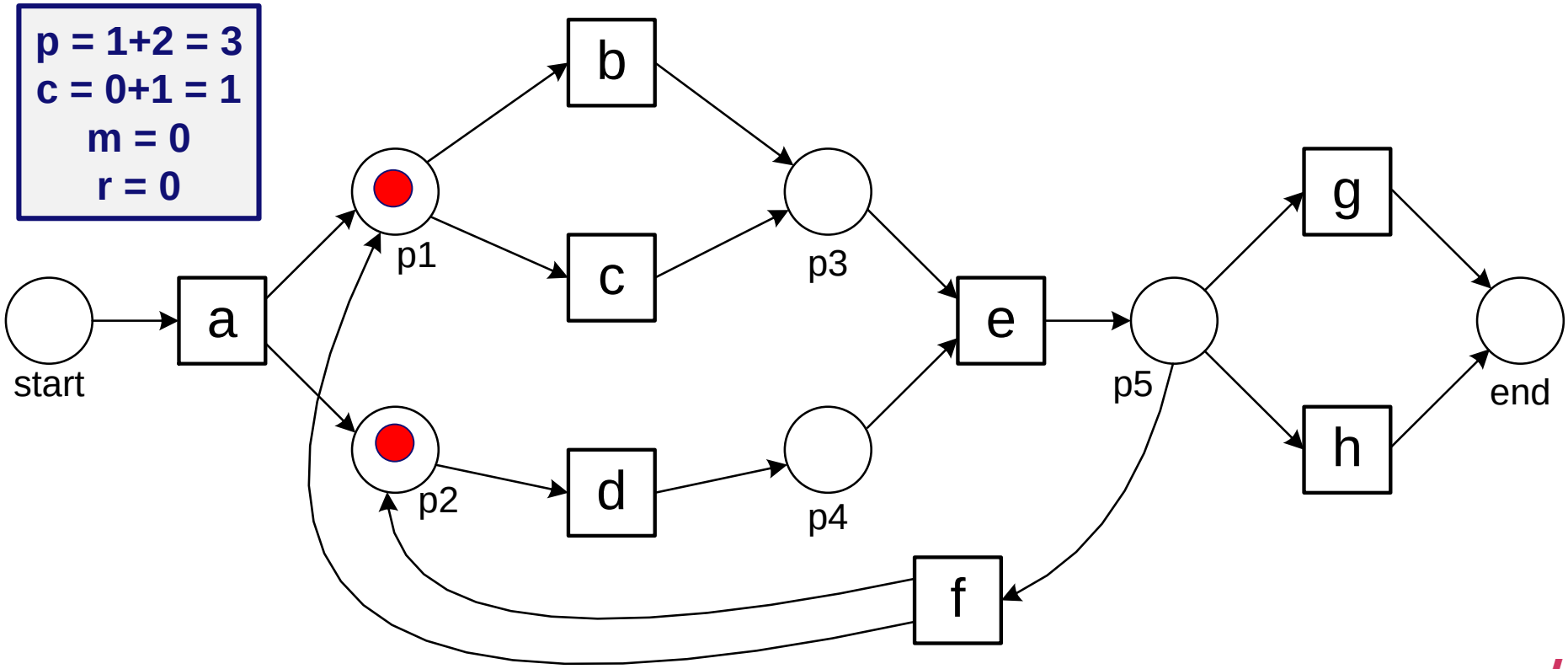
Replaying

$$\sigma_1 = \langle a\, c, d, e, h \rangle$$



TU/e

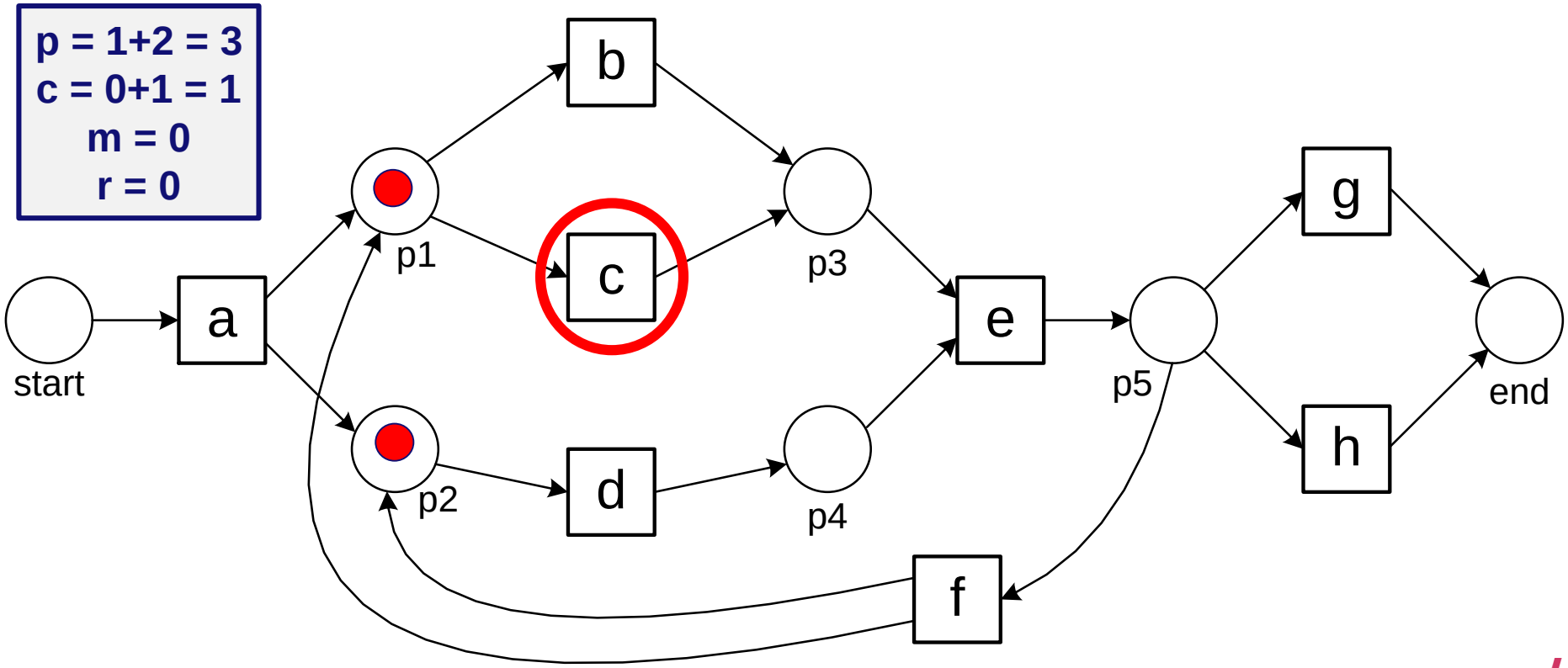
Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



TU/e

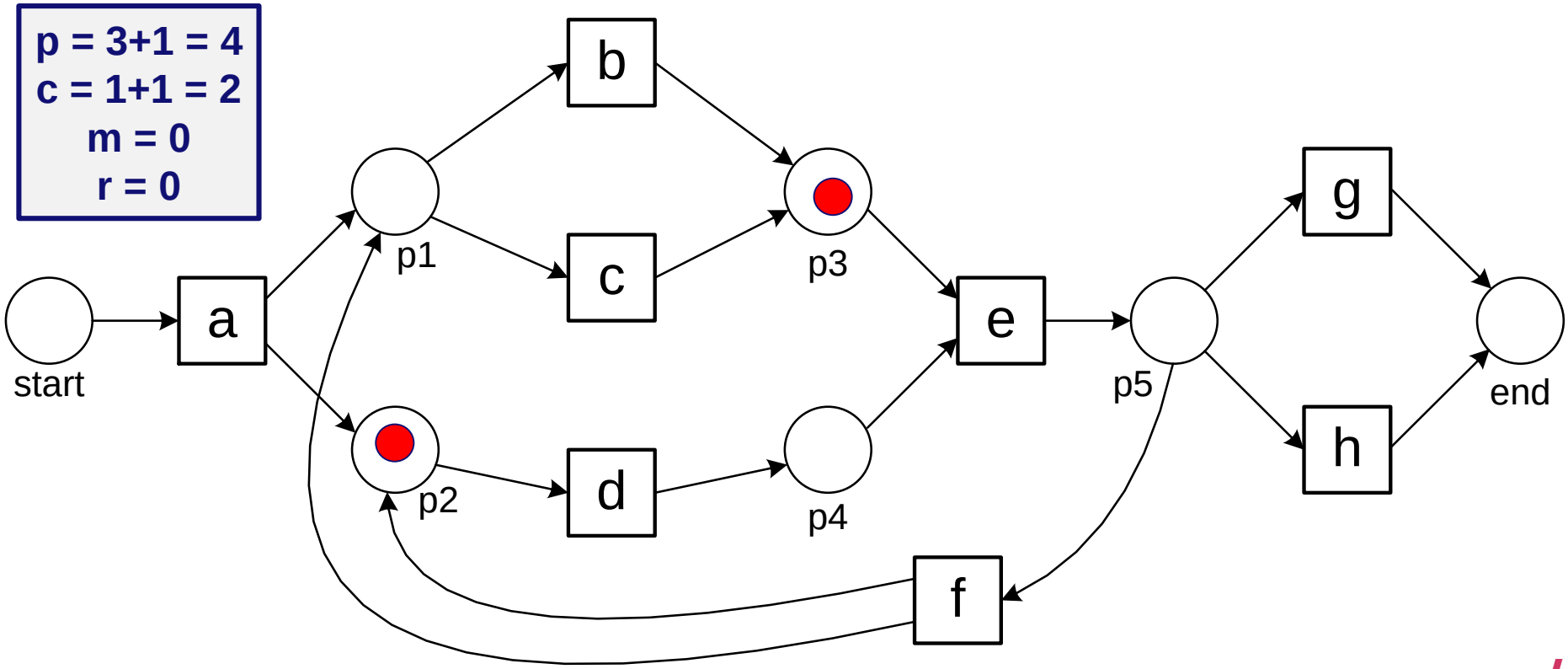
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



TU/e

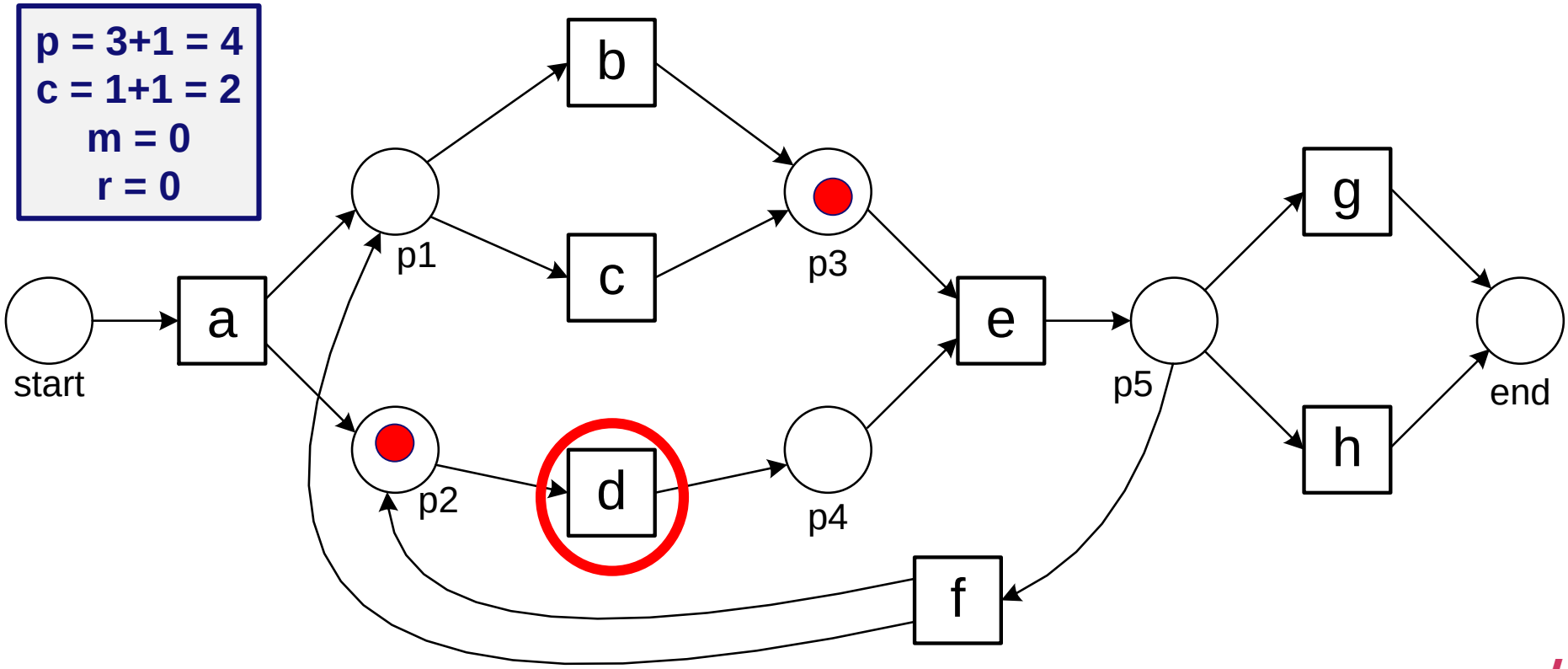
Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



TU/e

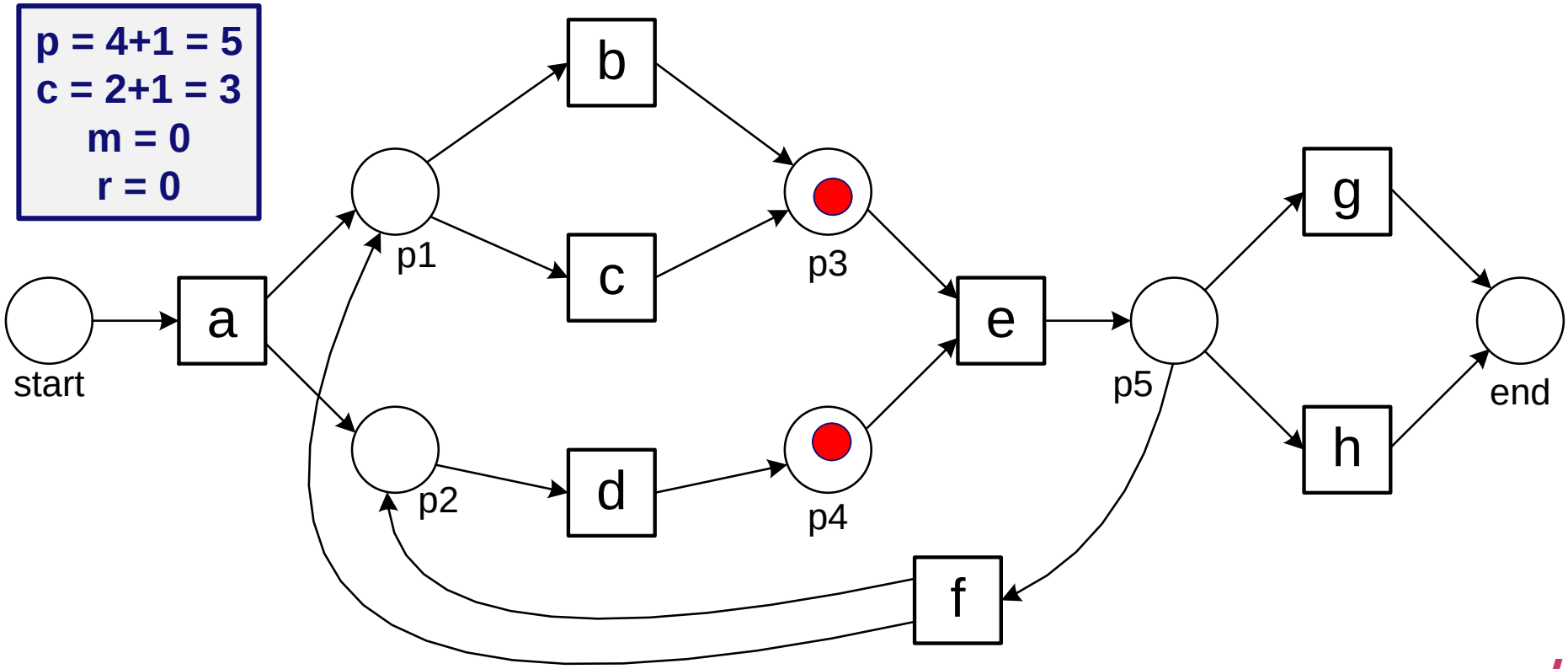
Replaying

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$

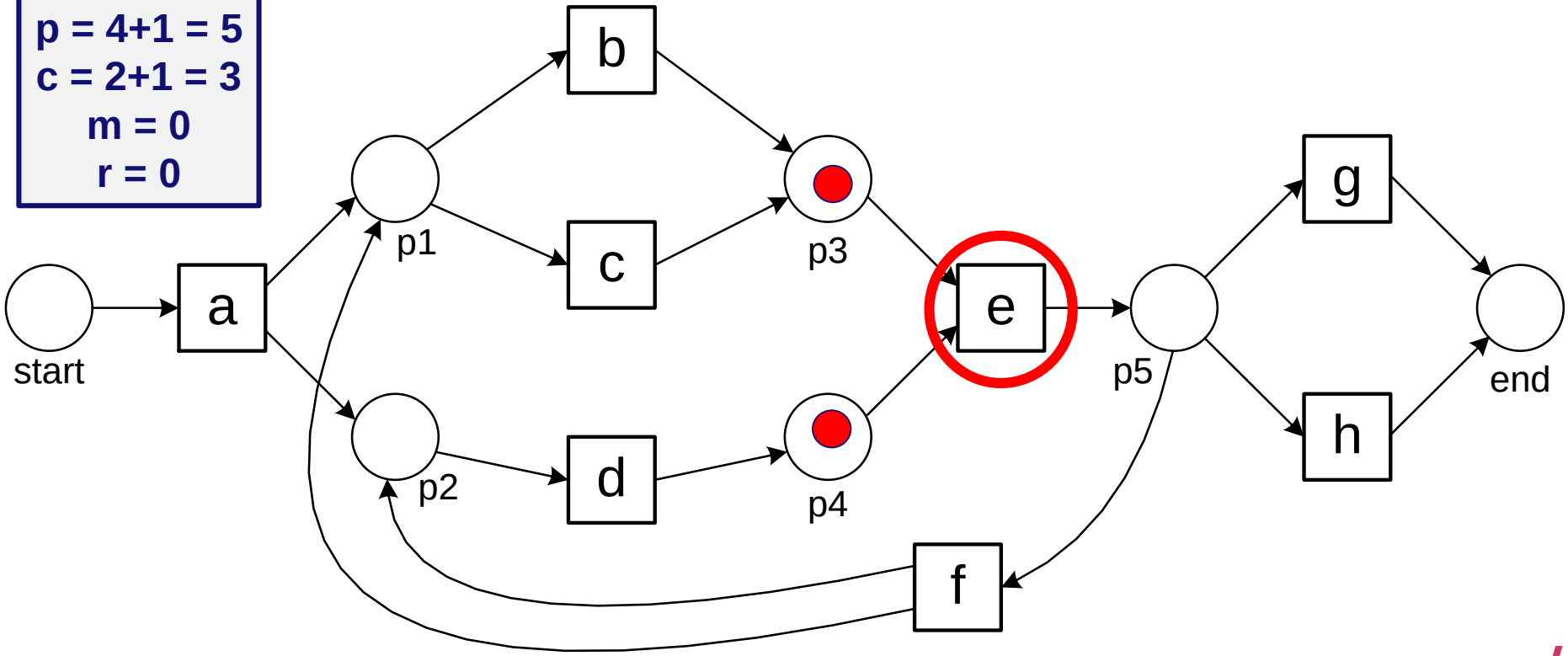


TU/e

Replaying

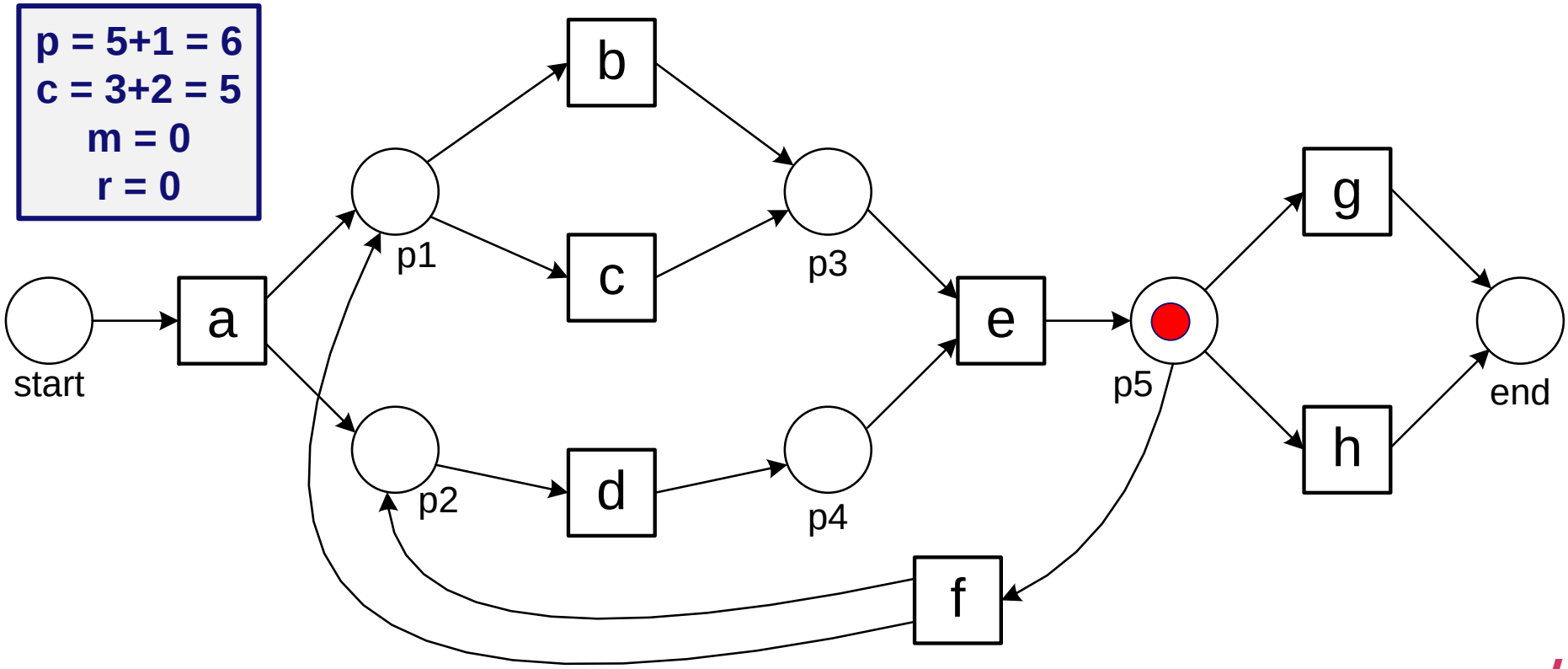
$$\sigma_1 = \langle a, c, d, e, h \rangle$$

p = 4+1 = 5
c = 2+1 = 3
m = 0
r = 0



TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$

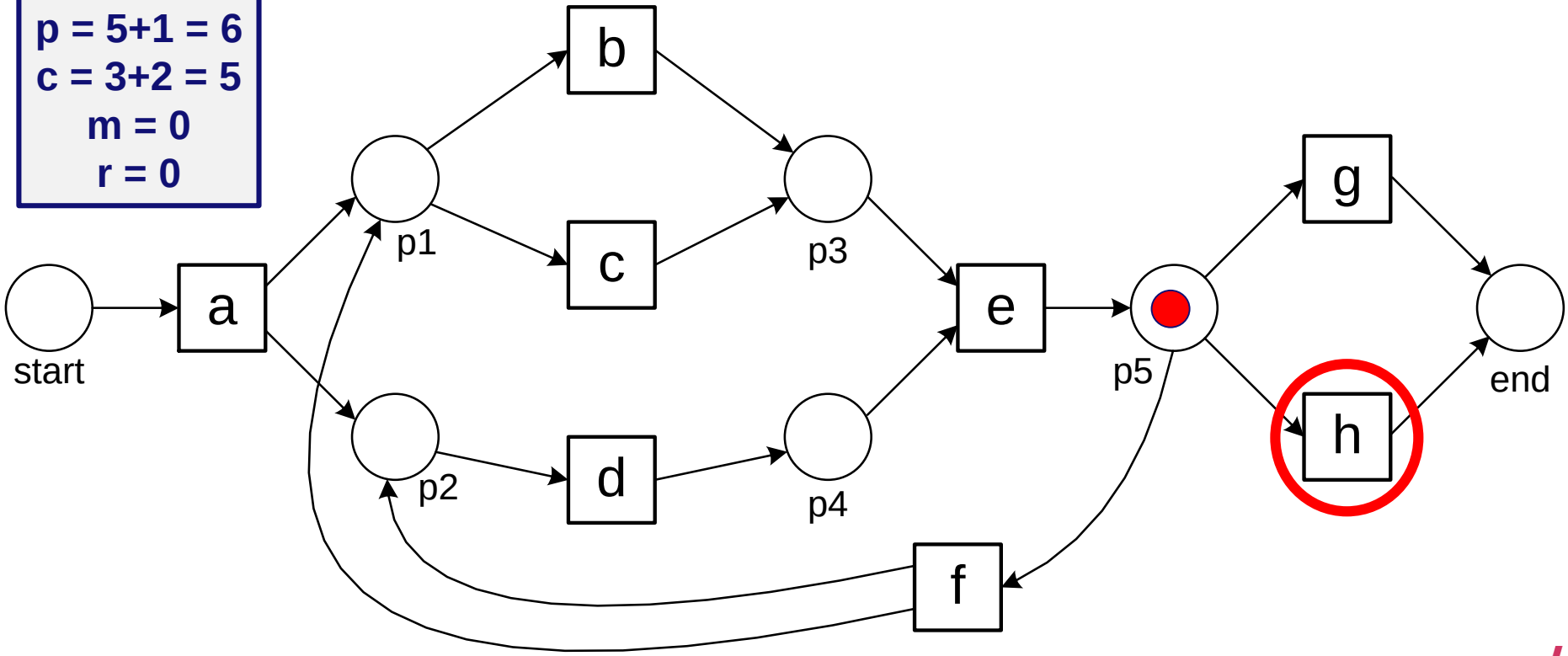


TU/e

Replaying

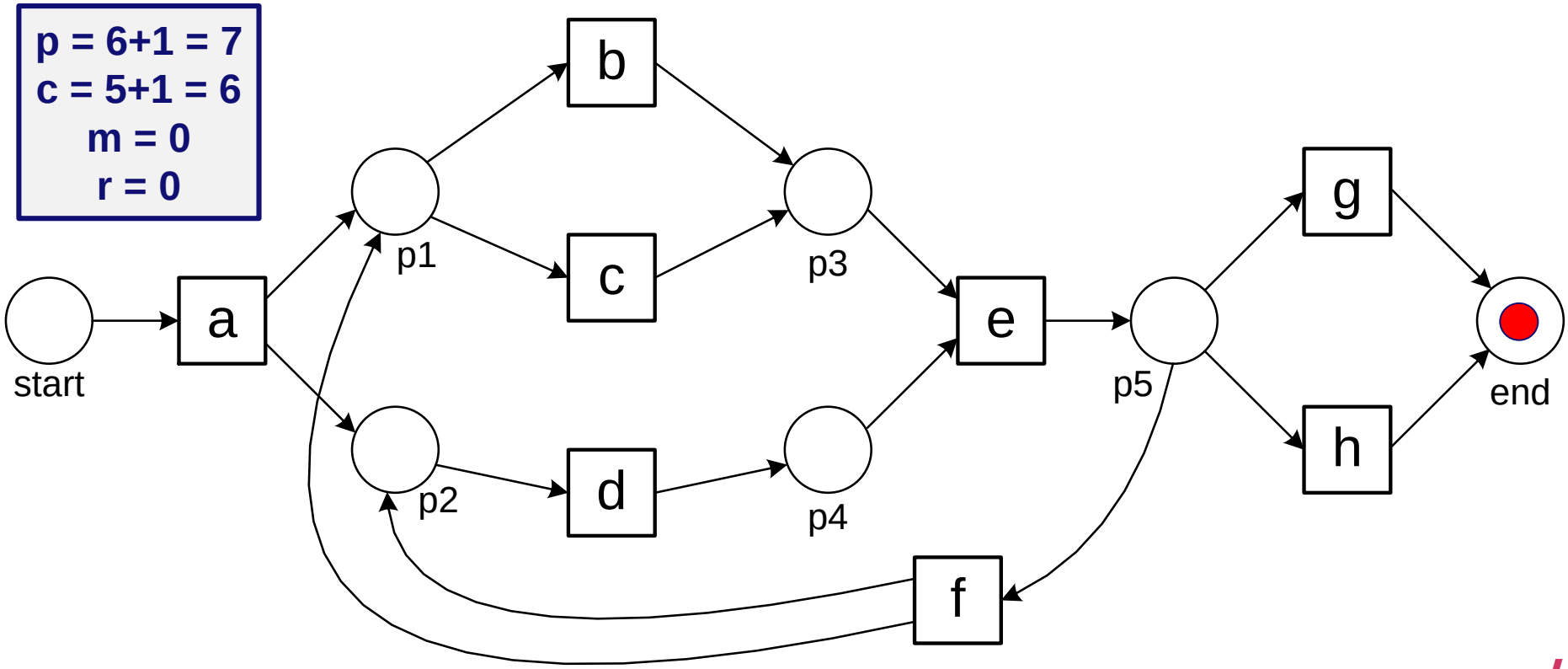
$$\sigma_1 = \langle a, c, d, e, h \rangle$$

p = 5+1 = 6
c = 3+2 = 5
m = 0
r = 0



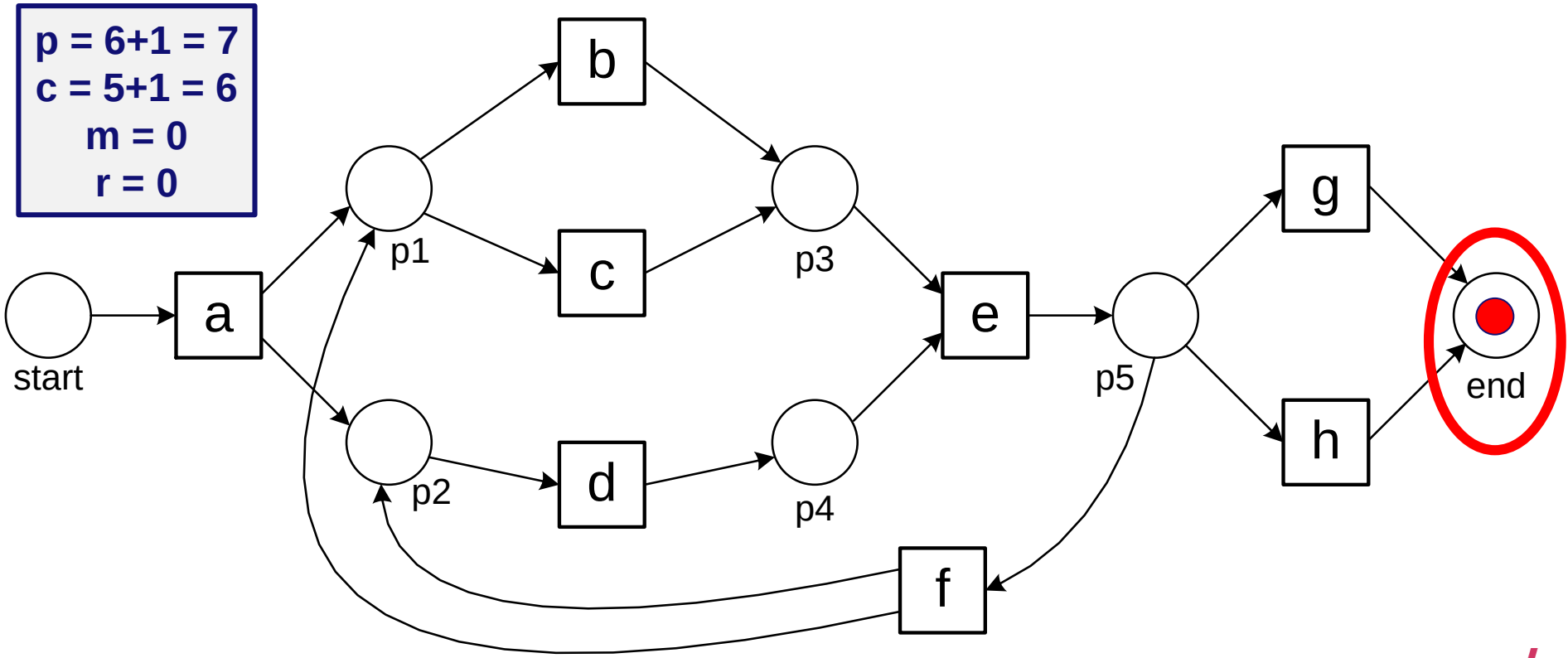
TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



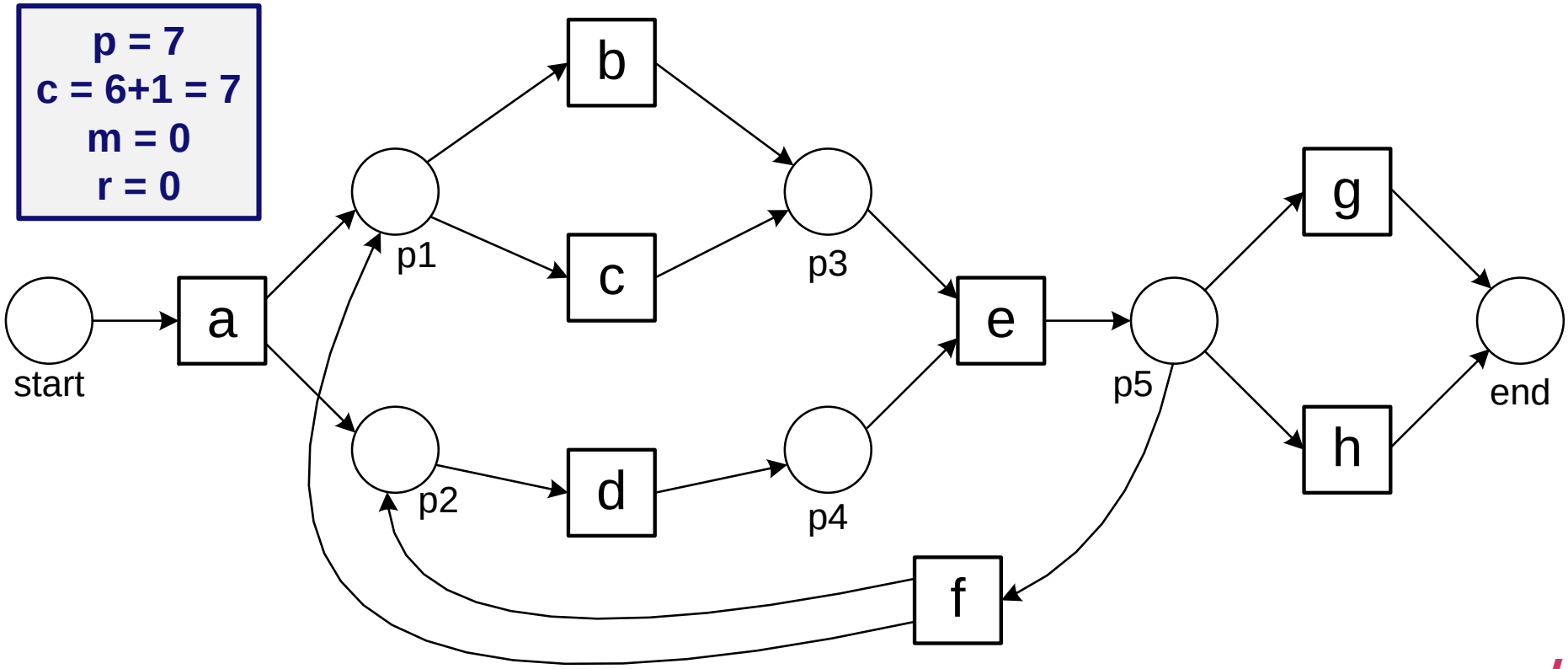
TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$



TU/e

Replaying $\sigma_1 = \langle a, c, d, e, h \rangle$

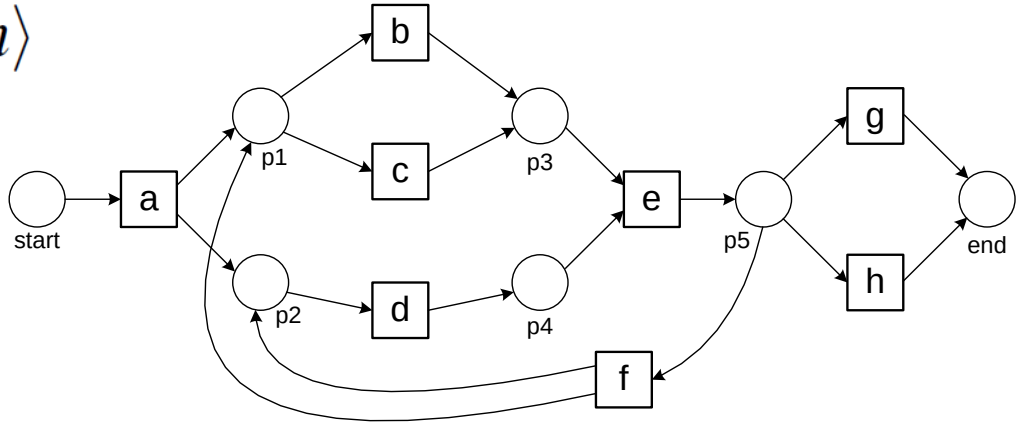


TU/e

Quantifying fitness at the trace level

p = 7
c = 7
m = 0
r = 0

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



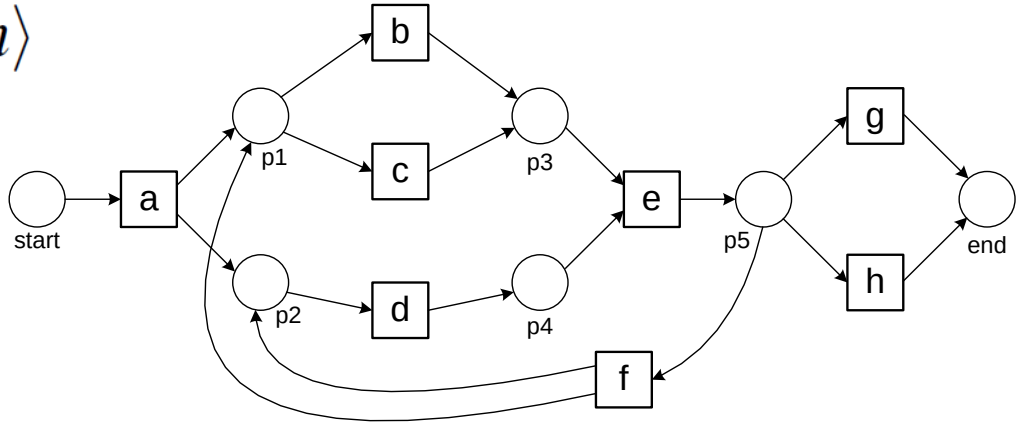
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

TU/e

Quantifying fitness at the trace level

p = 7
c = 7
m = 0
r = 0

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



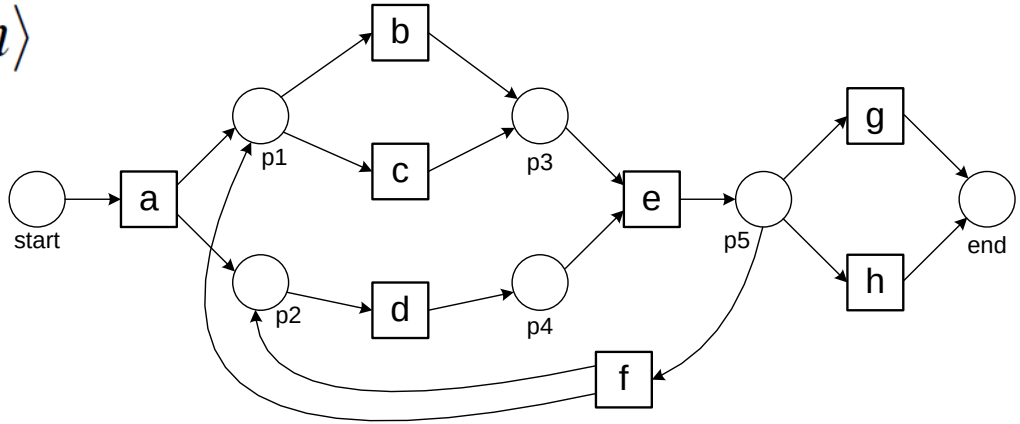
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{0}{7}\right) + \frac{1}{2}\left(1 - \frac{0}{7}\right)$$

TU/e

Quantifying fitness at the trace level

p = 7
c = 7
m = 0
r = 0

$$\sigma_1 = \langle a, c, d, e, h \rangle$$



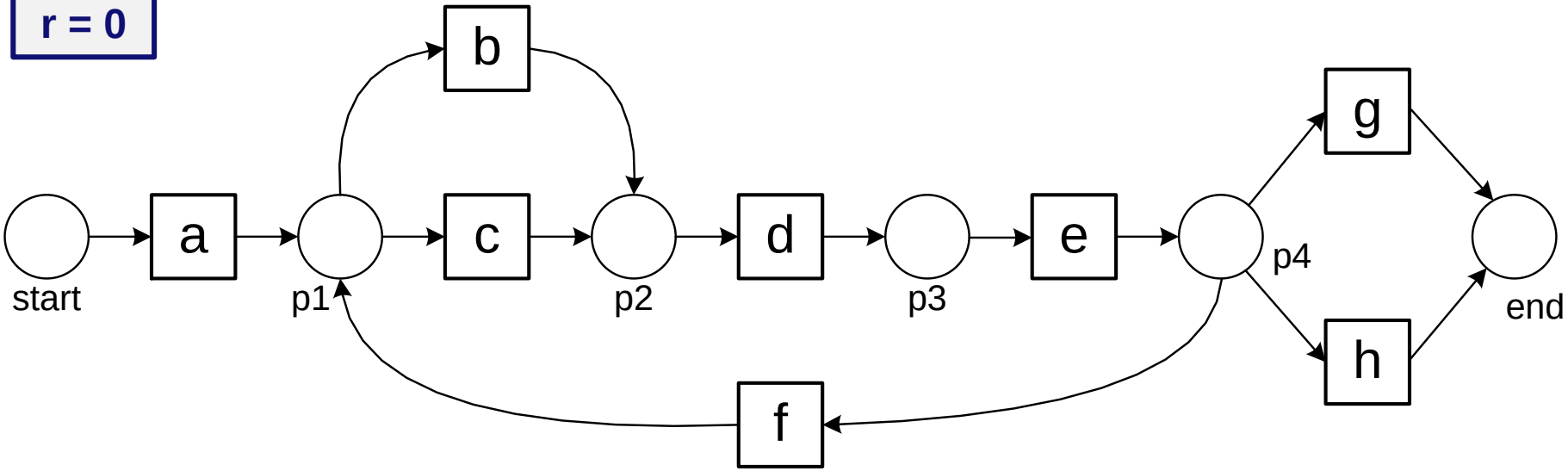
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{0}{7}\right) + \frac{1}{2}\left(1 - \frac{0}{7}\right) = 1$$

TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 0
c = 0
m = 0
r = 0

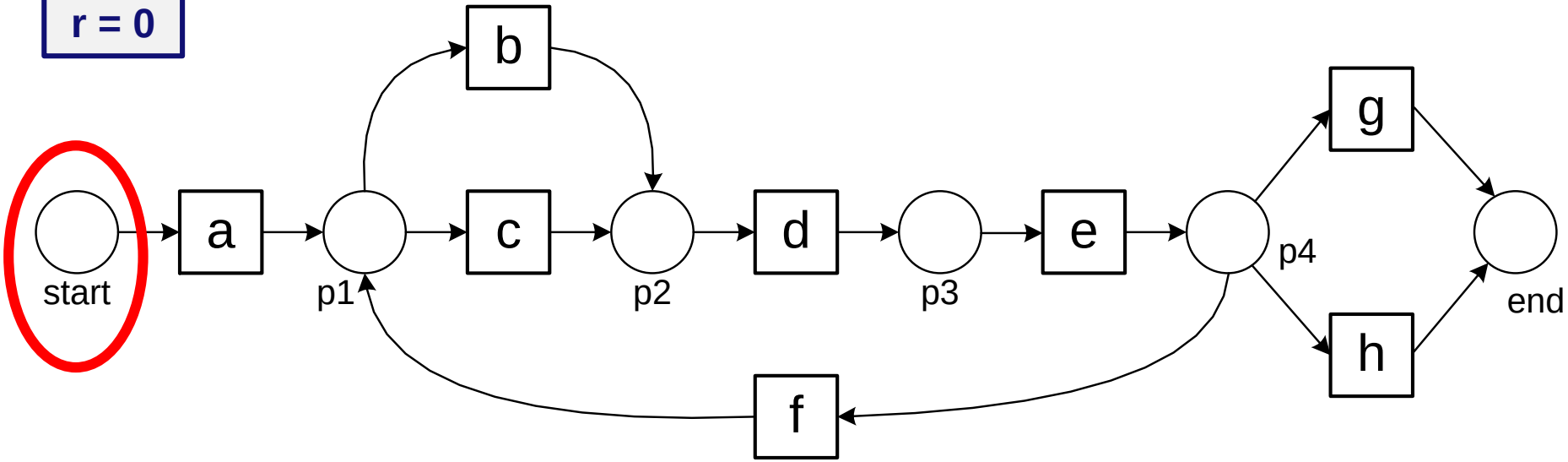


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 0
c = 0
m = 0
r = 0

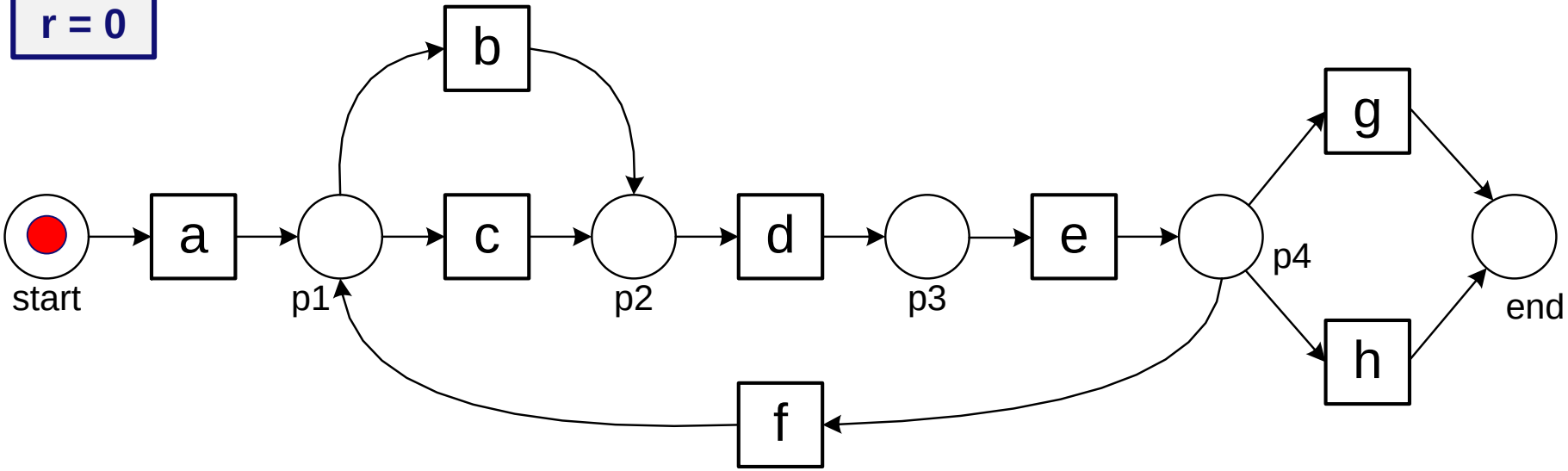


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 1
c = 0
m = 0
r = 0

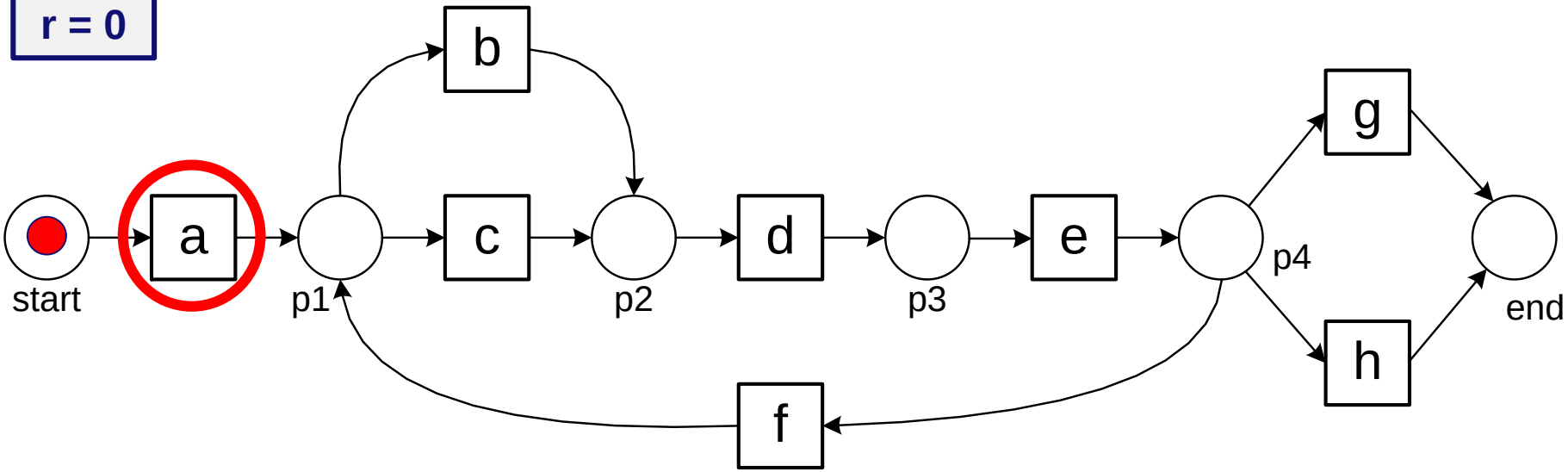


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 1
c = 0
m = 0
r = 0

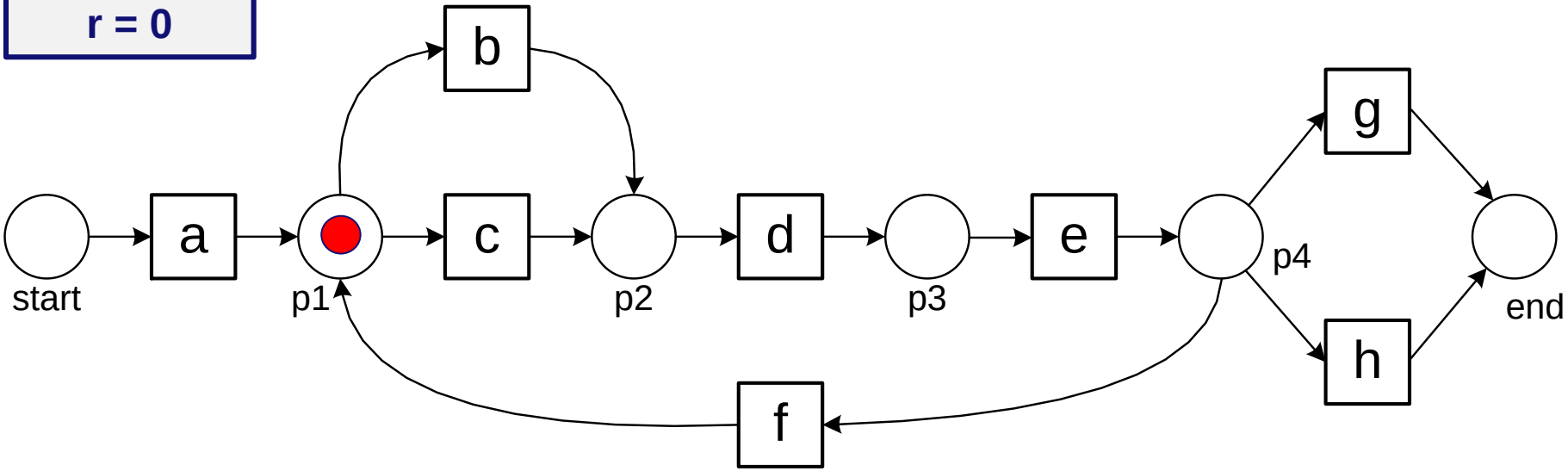


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 1+1 = 2
c = 0+1 = 1
m = 0
r = 0

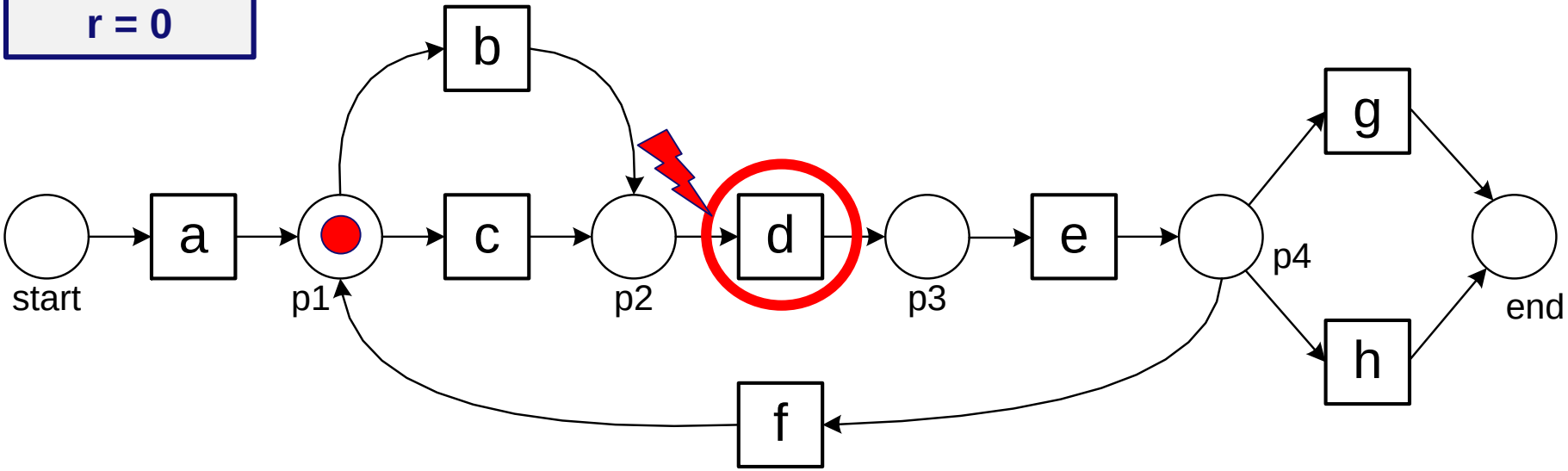


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 1+1 = 2
c = 0+1 = 1
m = 0
r = 0

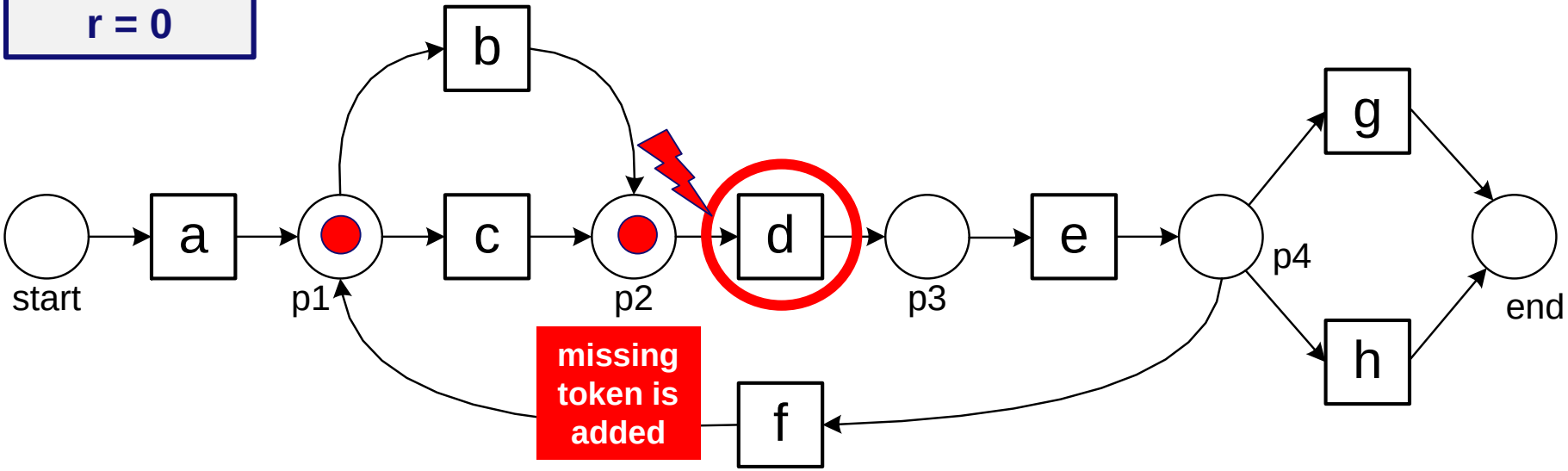


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 1+1 = 2
c = 0+1 = 1
m = 0
r = 0

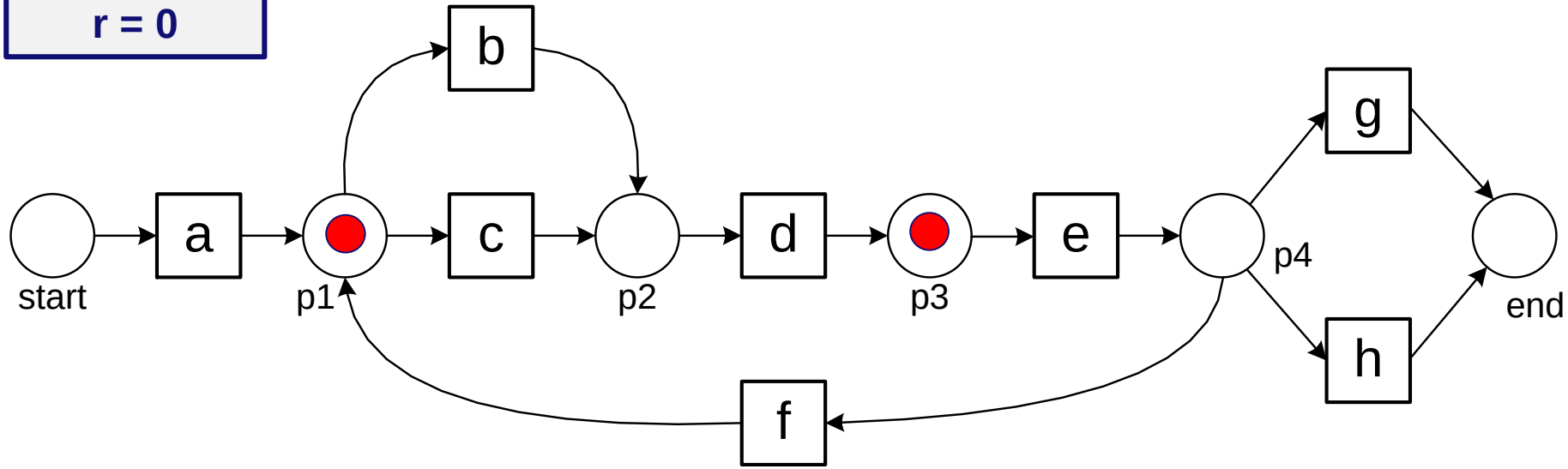


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 2+1 = 3
c = 1+1 = 2
m = 0+1 = 1
r = 0

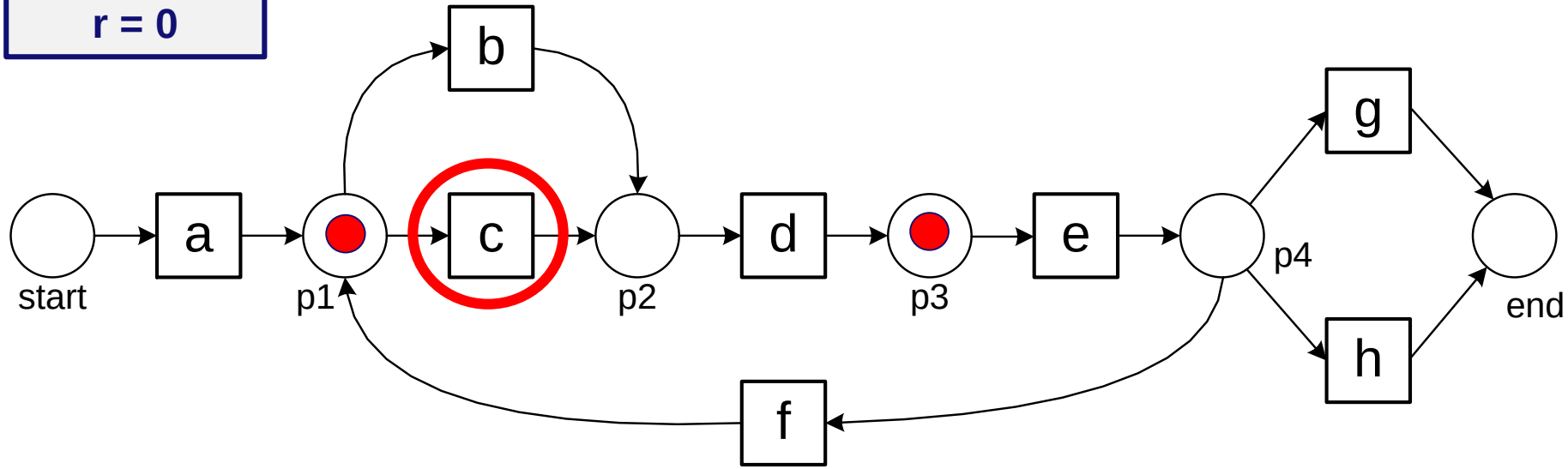


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 2+1 = 3
c = 1+1 = 2
m = 0+1 = 1
r = 0

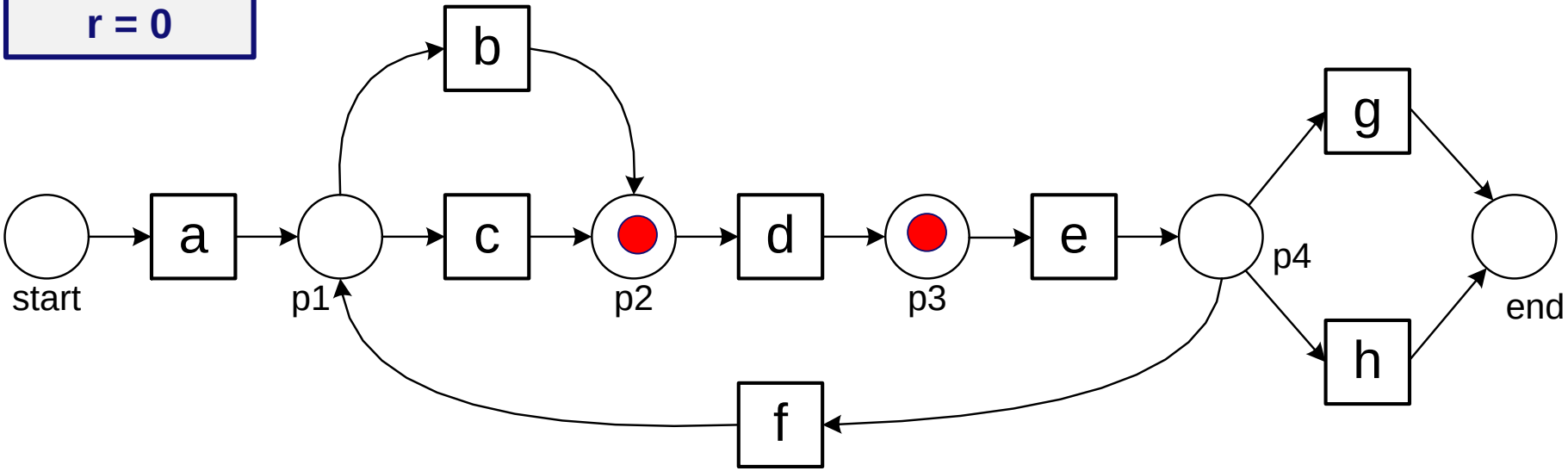


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 3+1 = 4
c = 2+1 = 3
m = 1
r = 0

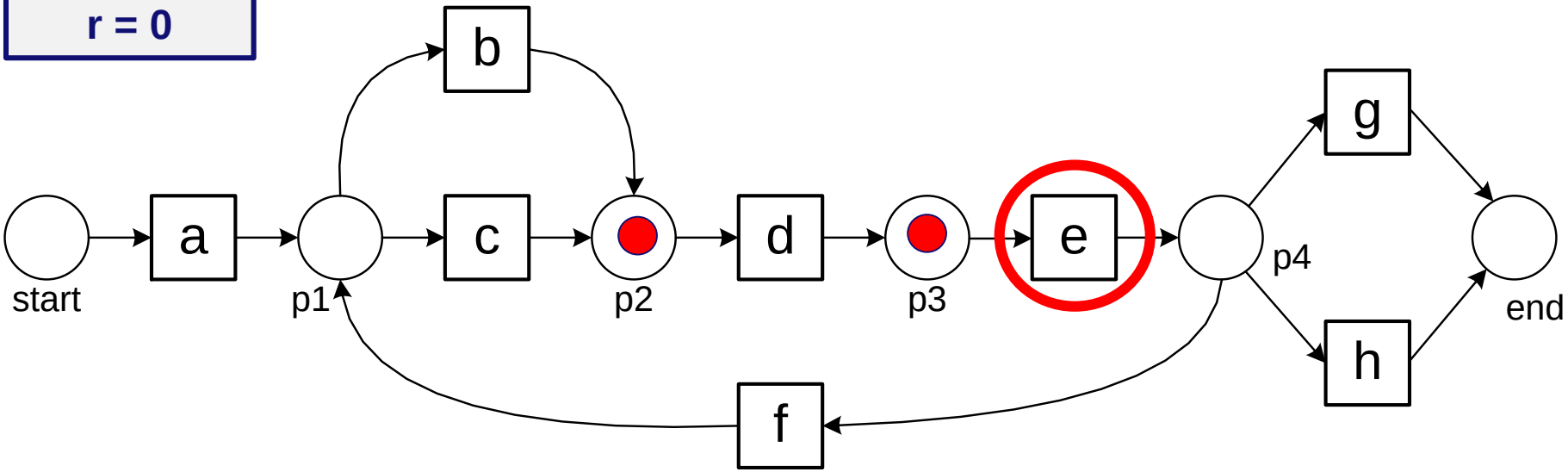


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 3+1 = 4
c = 2+1 = 3
m = 1
r = 0

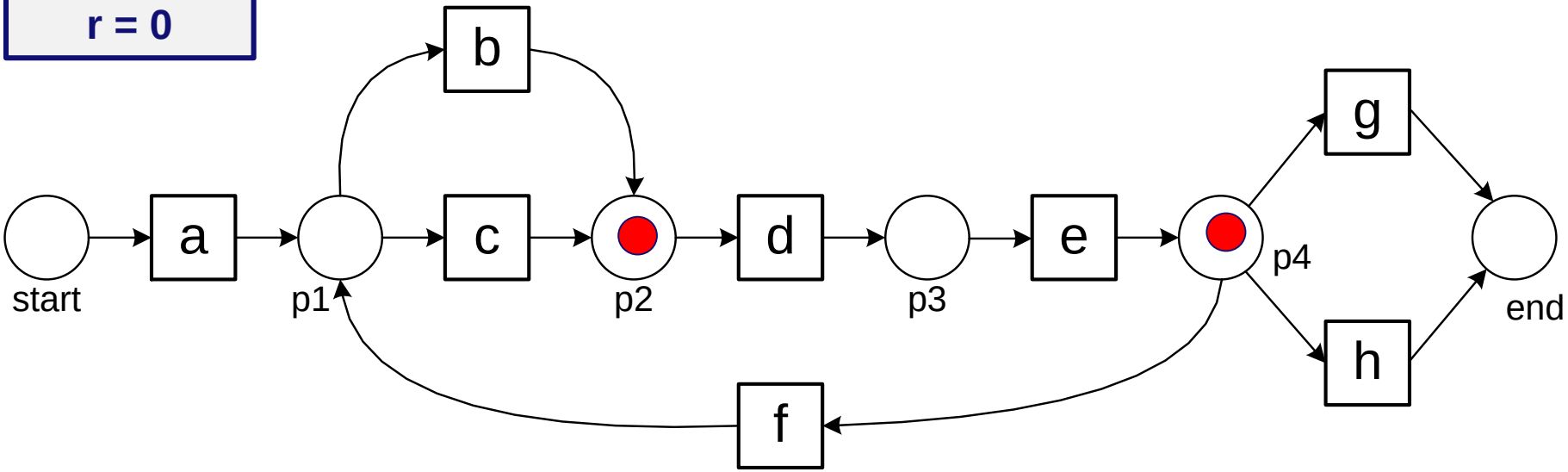


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 4+1 = 5
c = 3+1 = 4
m = 1
r = 0

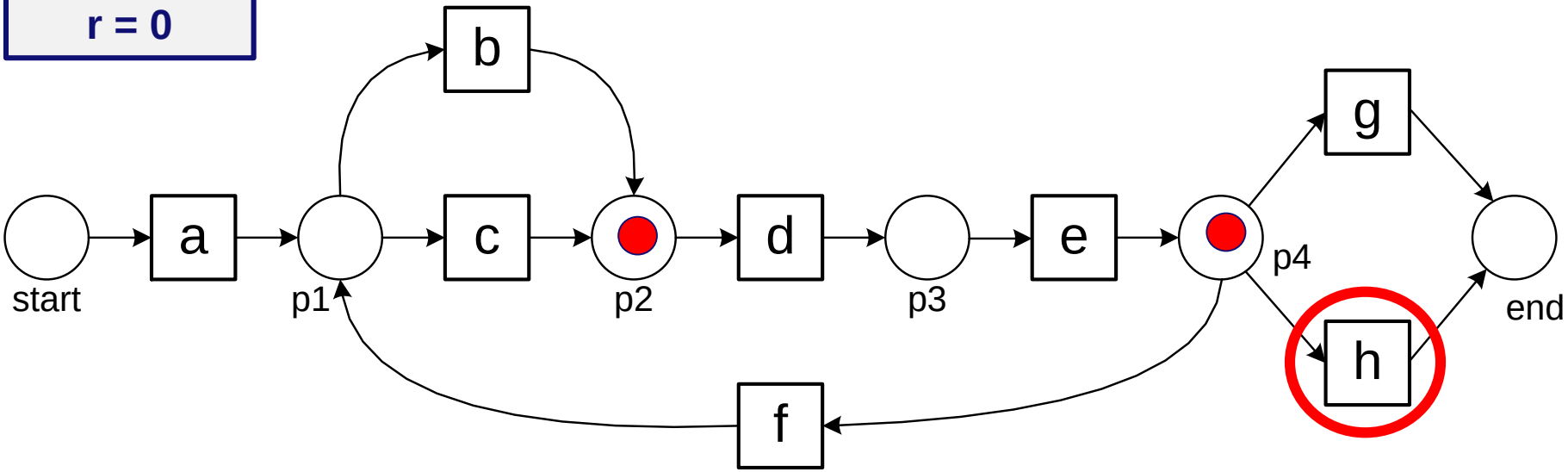


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 4+1 = 5
c = 3+1 = 4
m = 1
r = 0

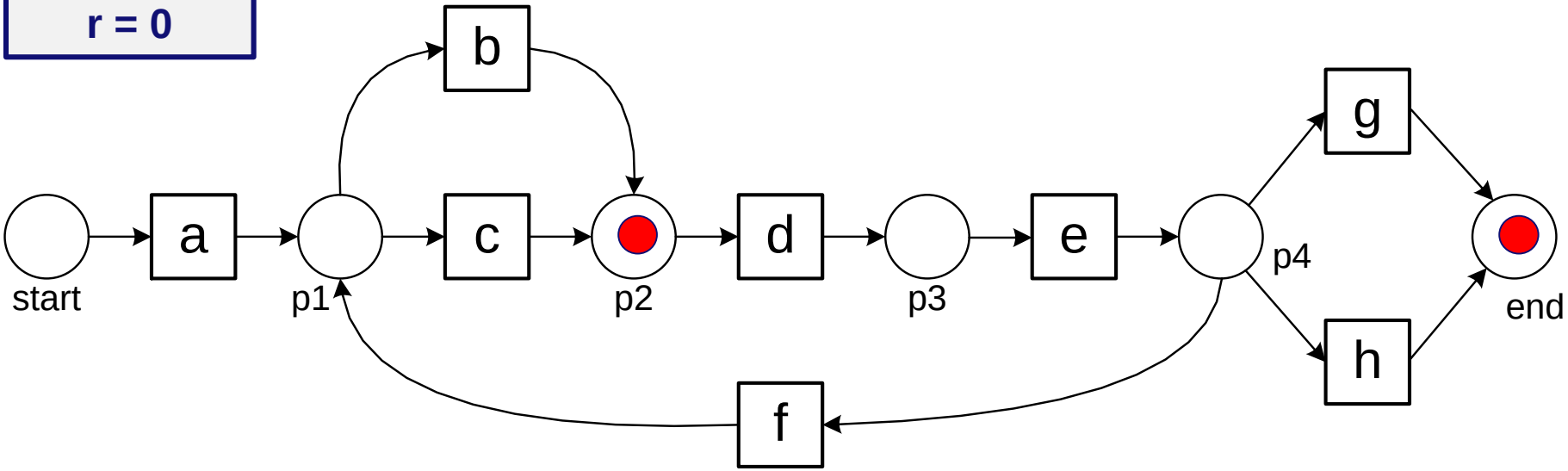


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 5+1 = 6
c = 4+1 = 5
m = 1
r = 0

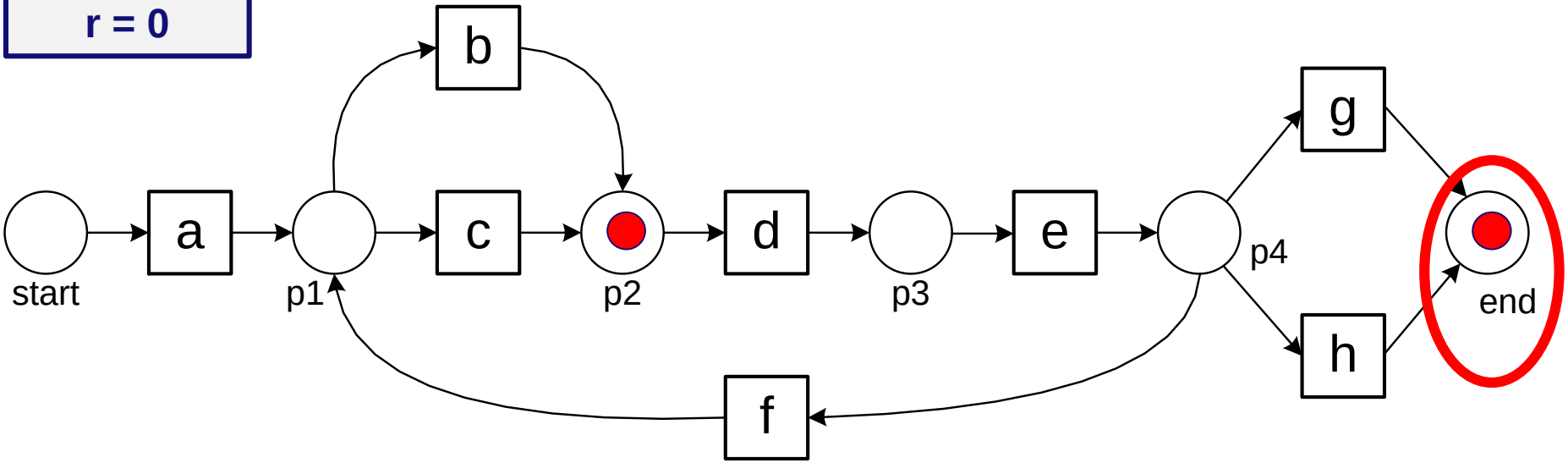


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 5+1 = 6
c = 4+1 = 5
m = 1
r = 0

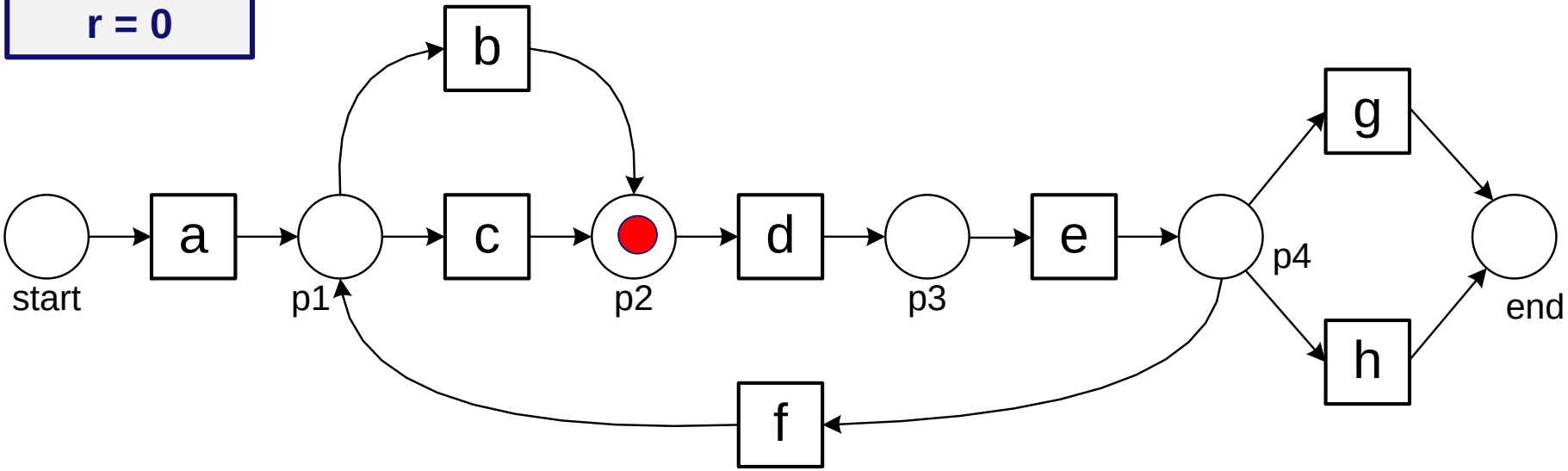


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 6
c = 5+1 = 6
m = 1
r = 0

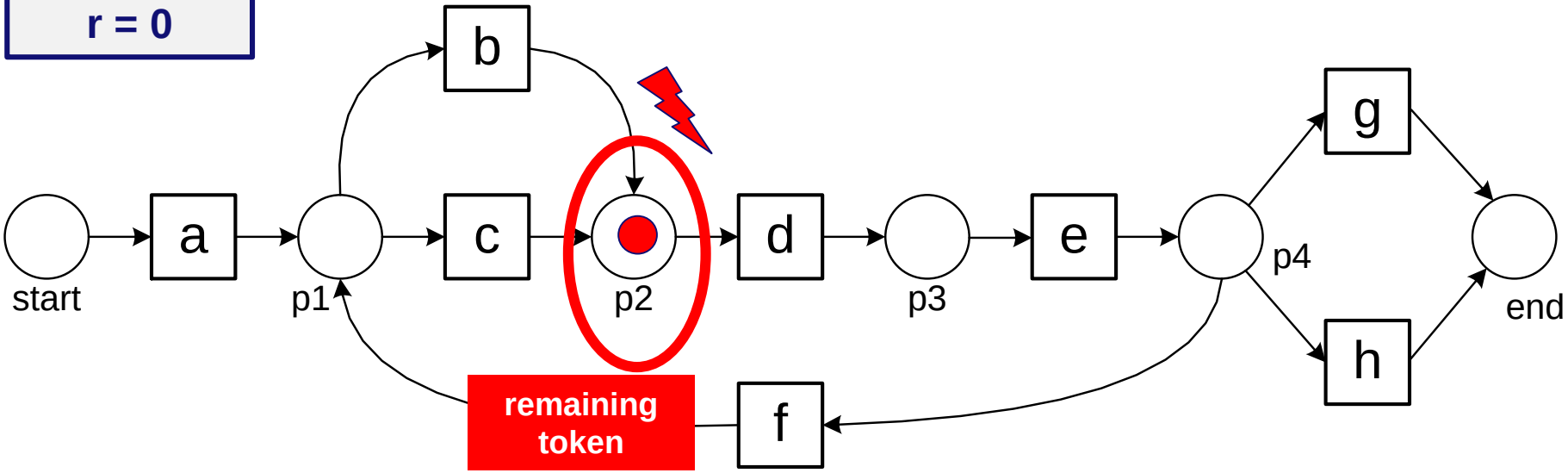


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 6
c = 5+1 = 6
m = 1
r = 0

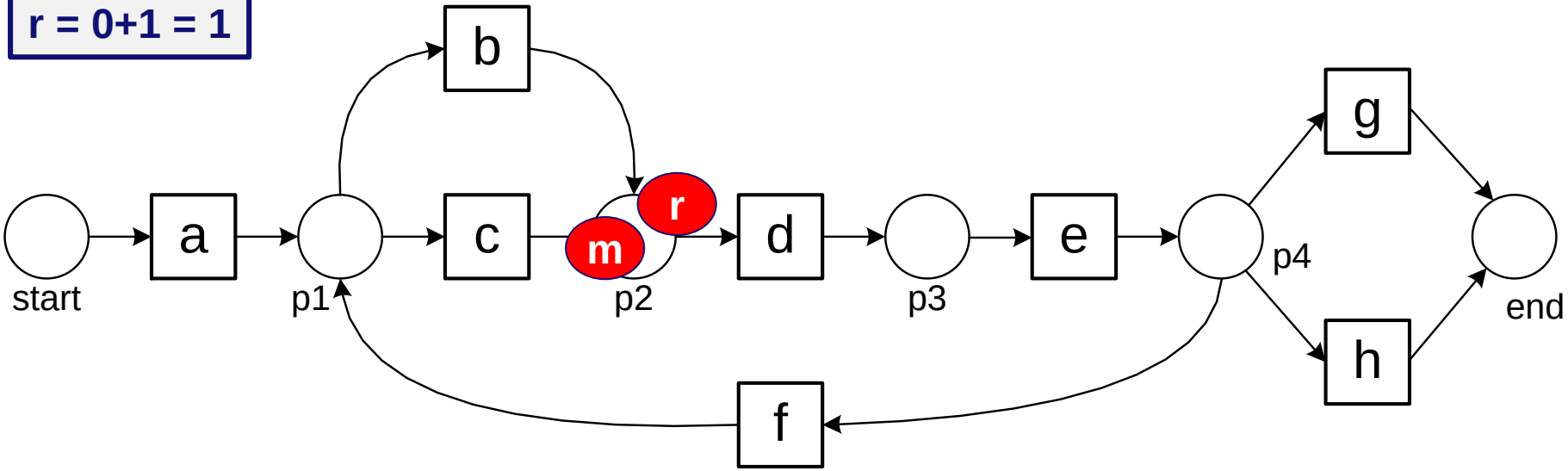


TU/e

Replaying

$$\sigma_3 = \langle a, d, c, e, h \rangle$$

p = 6
c = 6
m = 1
r = 0+1 = 1

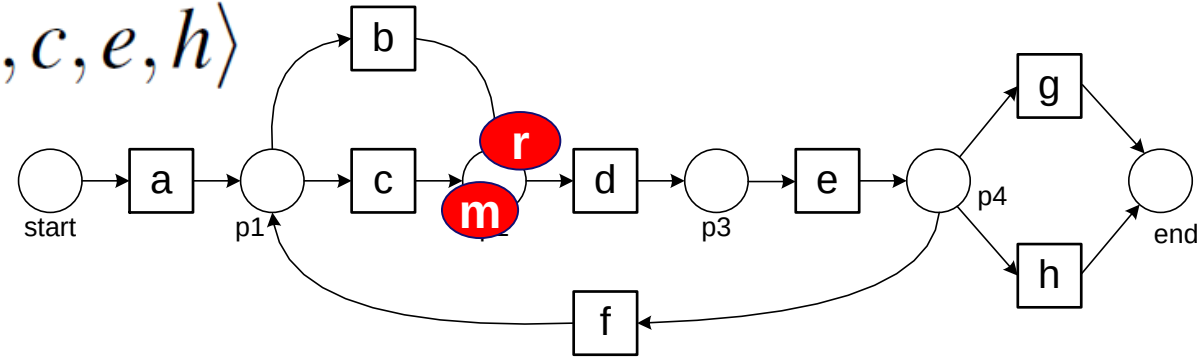


TU/e

Quantifying fitness at the trace level

p = 6
c = 6
m = 1
r = 1

$$\sigma_3 = \langle a, d, c, e, h \rangle$$



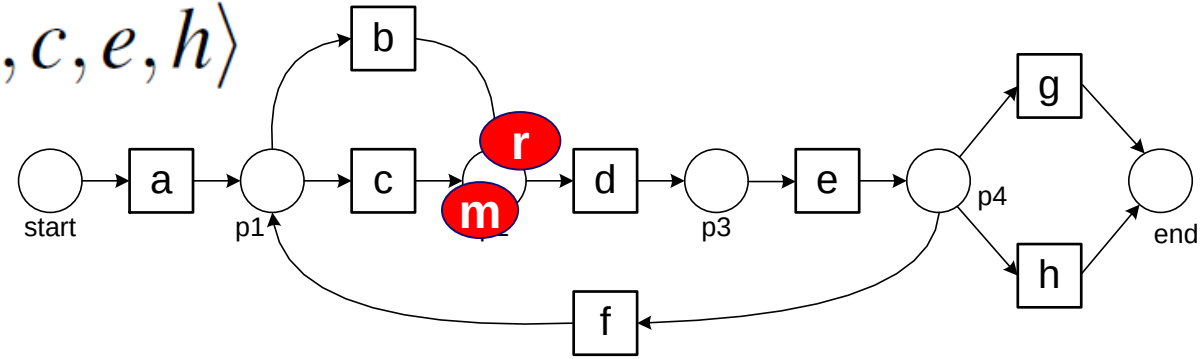
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

TU/e

Quantifying fitness at the trace level

p = 6
c = 6
m = 1
r = 1

$$\sigma_3 = \langle a, d, c, e, h \rangle$$



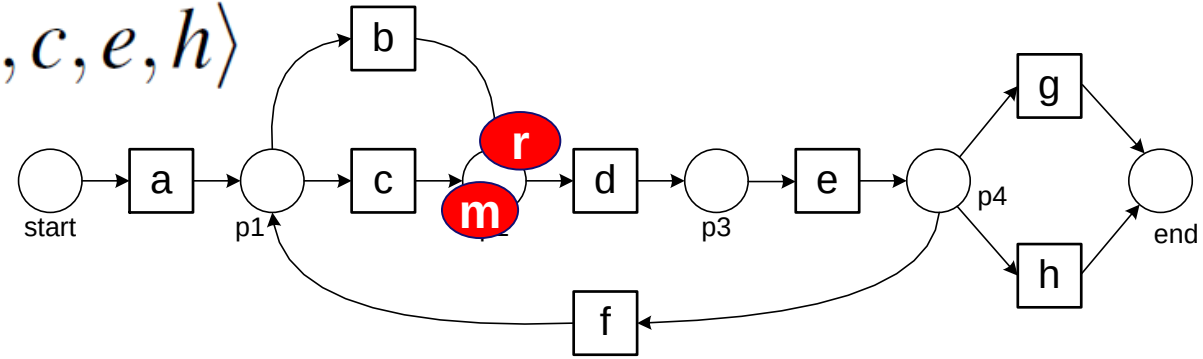
$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right)$$

TU/e

Quantifying fitness at the trace level

p = 6
c = 6
m = 1
r = 1

$$\sigma_3 = \langle a, d, c, e, h \rangle$$



$$fitness(\sigma, N) = \frac{1}{2}\left(1 - \frac{1}{6}\right) + \frac{1}{2}\left(1 - \frac{1}{6}\right) = 0.8333$$

TU/e

Fitness at the log level

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

TU/e

Fitness at the log level

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

Looks scary, but one just needs to take the sums of p, c, m, and r over the multiset of traces in de event log …

TU/e

Fitness at the log level

missing tokens

consumed tokens

remaining tokens

produced tokens

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

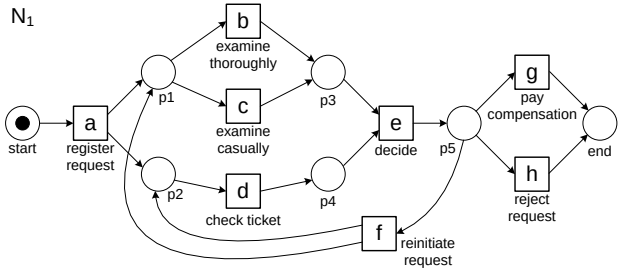
Looks scar
just needs t
sums of p, c, m, and r
over the multiset of
traces in de

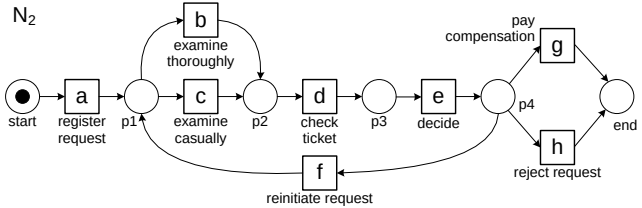
TU/e

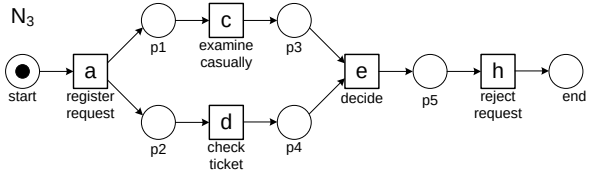
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

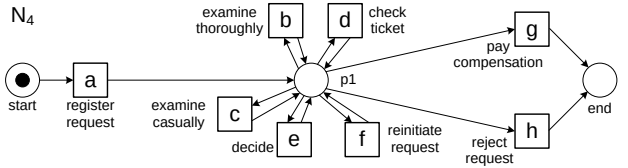
?

$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$









#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

?

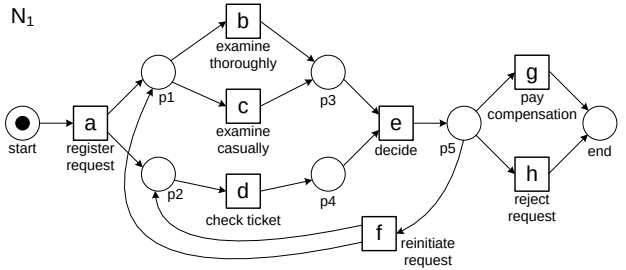
$$fitness(L,N) = \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times m_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times c_{N,\sigma}}\right) + \frac{1}{2}\left(1 - \frac{\sum_{\sigma \in L} L(\sigma) \times r_{N,\sigma}}{\sum_{\sigma \in L} L(\sigma) \times p_{N,\sigma}}\right)$$

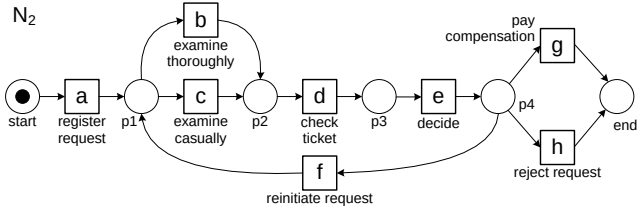
$$fitness(L_{full}, N_1) = 1$$

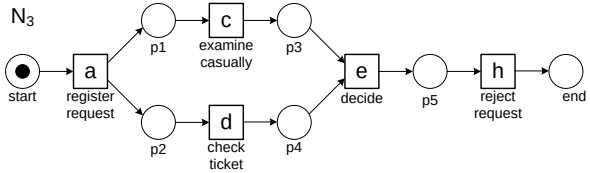
$$fitness(L_{full}, N_2) = 0.9504$$

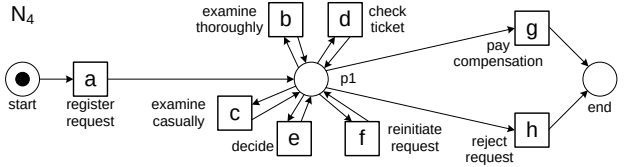
$$fitness(L_{full}, N_3) = 0.8797$$

$$fitness(L_{full}, N_4) = 1$$

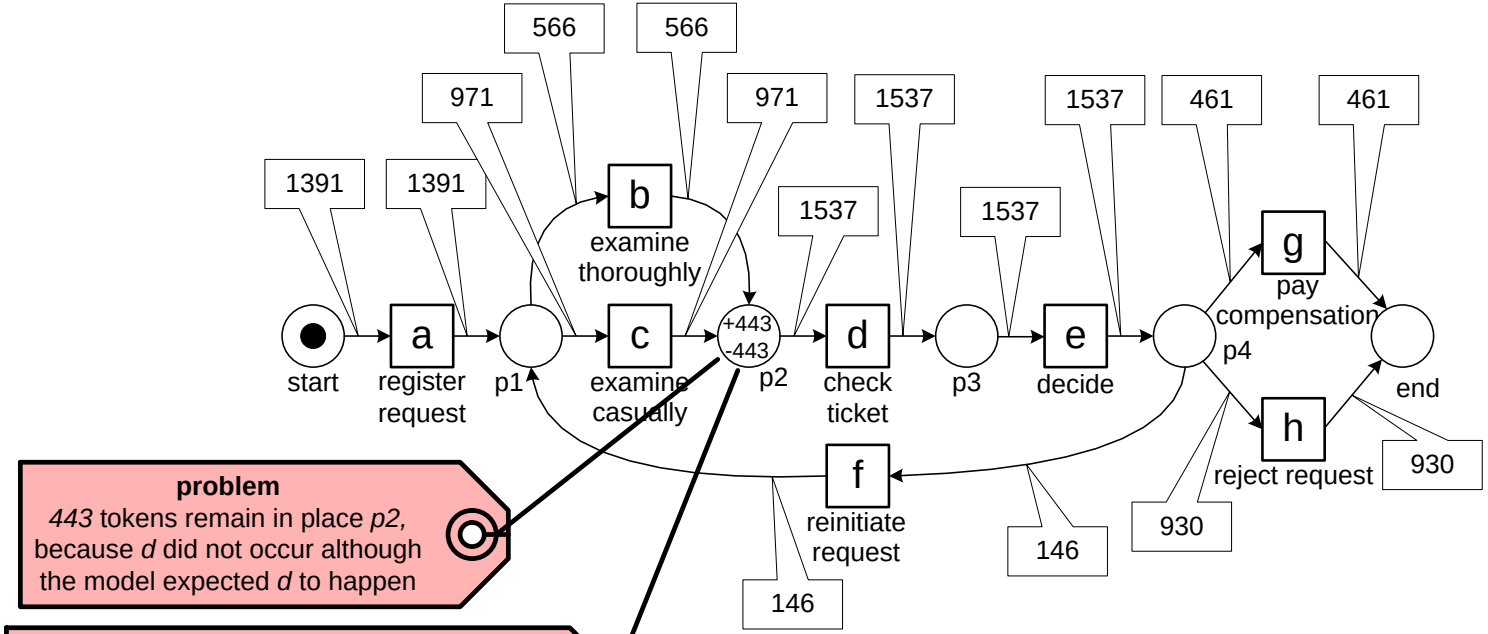








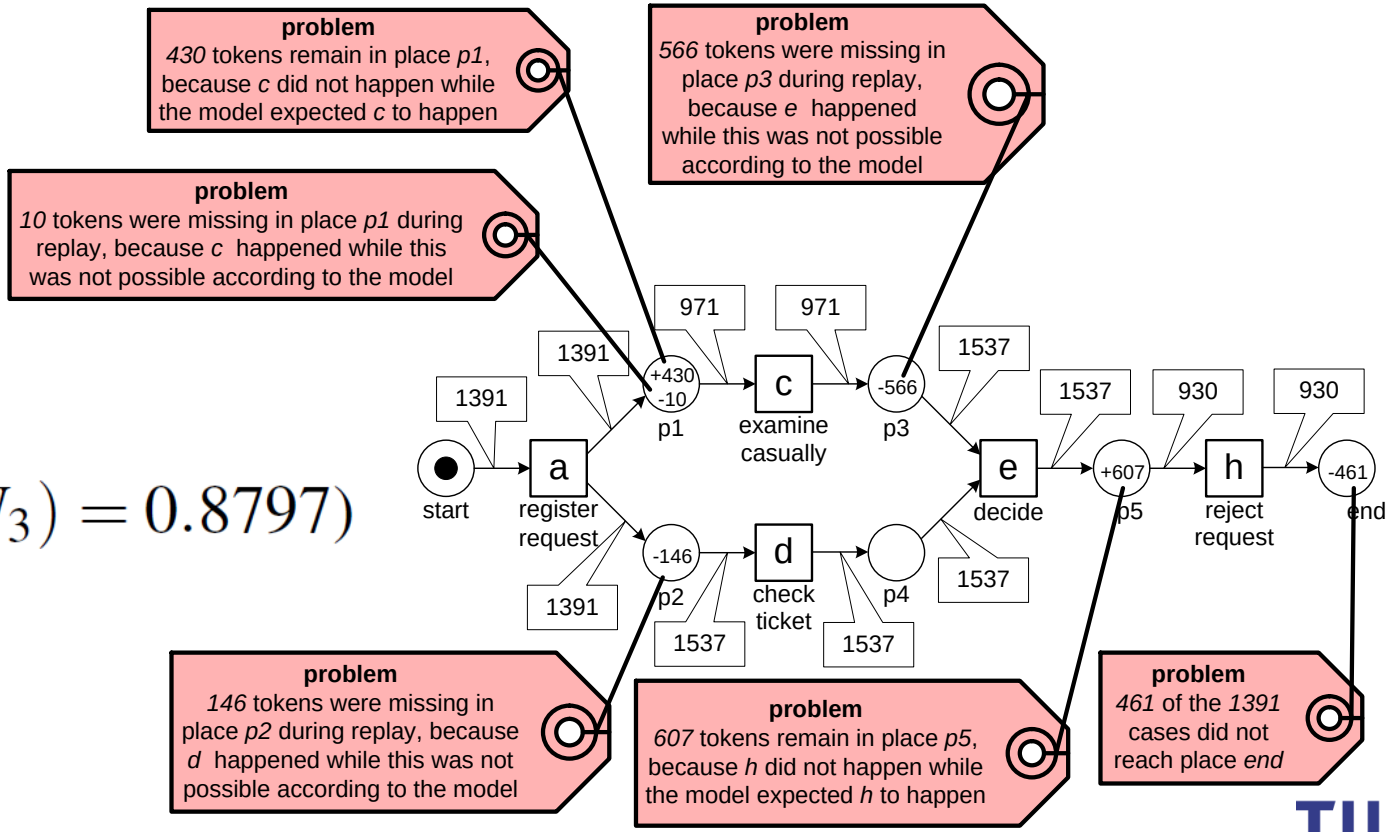
Diagnostics



$$(fitness(L_{full}, N_2) = 0.9504)$$

Diagnostics

$$(fitness(L_{full}, N_3) = 0.8797)$$



Part I: Introduction

Chapter 1
Data Science in Action

Chapter 2
Process Mining: The Missing Link

Part II: Preliminaries

Chapter 3
Process Modeling and Analysis

Chapter 4
Data Mining

Part III: From Event Logs to Process Models

Chapter 5
Getting the Data

Chapter 6
Process Discovery: An Introduction

Chapter 7
Advanced Process Discovery Techniques

Part IV: Beyond Process Discovery

Chapter 8
Conformance Checking

Chapter 9
Mining Additional Perspectives

Chapter 10
Operational Support

Part V: Putting Process Mining to Work

Chapter 11
Process Mining Software

Chapter 12
Process Mining in the Large

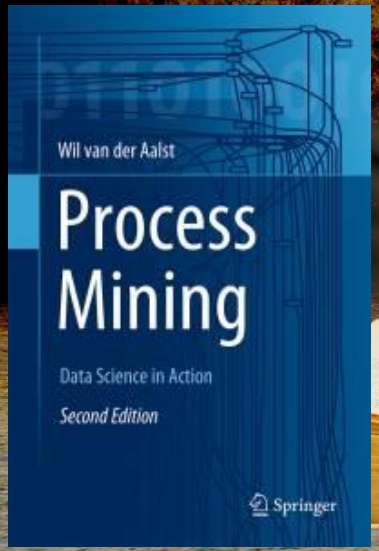
Chapter 13
Analyzing "Lasagna Processes"

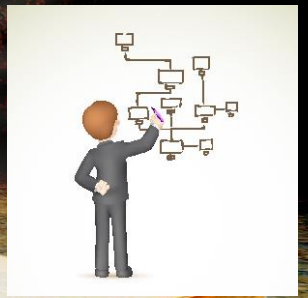
Chapter 14
Analyzing "Spaghetti Processes"

Part VI: Reflection

Chapter 15
Cartography and Navigation

Chapter 16
Epilogue





TU/e