Tick..tick..tick.. BOOM!

How to size thread pools and timeouts without getting caught in an explosion

- Christian Rehn – JUG KA 2020-02-12

In case of an incident:
Increase or decrease timeouts?

Christian Rehn

- Senior Software Developer
- Blog: http://www.christian-rehn.de
- Design Knights: http://design-types.net/ – http://principles-wiki.net/

This presentation in written form

http://www.christian-rehn.de/2019/10/03/die-praesidentenmacherin-und-der-timeout/

Timeout = avg(responseTime) + 50%

100%

200%

Timeouts are there to Handle High-Load Scenarios

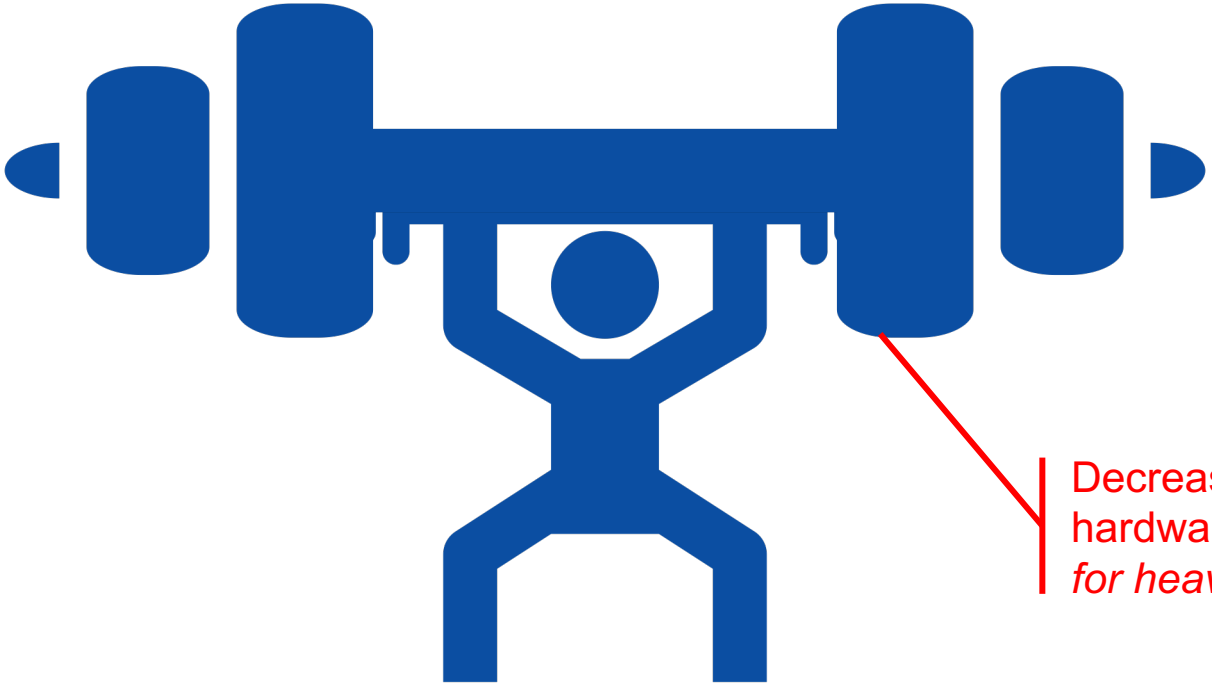
How long
am I willing to wait?

Some Rules of Thumb

Time	How users react
< 0,1s	“There is no delay. I’m operating on the data directly!”
0,1s … 1s	“There is some delay but I don’t have to wait.”
>1s	“I have to wait.”
~ 10s	“That’s long. I better do something else in the meanwhile…”
~ 15s, maybe 20s	Maximum tolerable waiting time

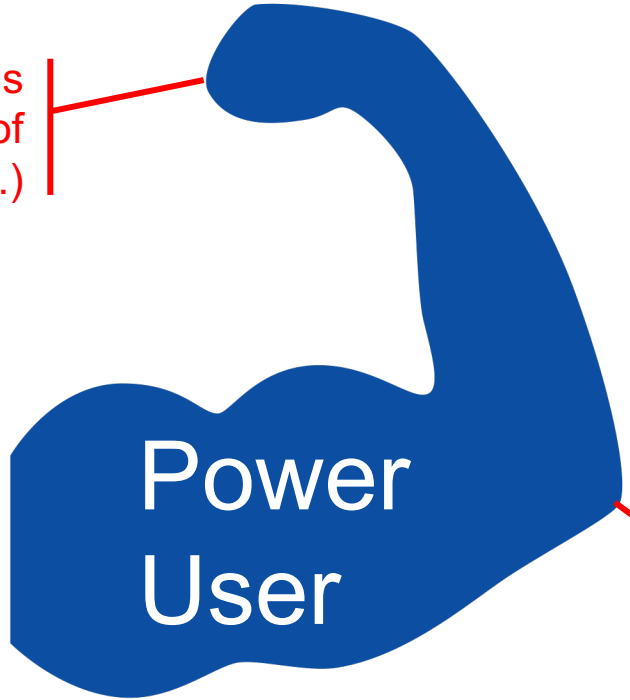
Here it is: a sensible value for a timeout!

http://sighci.org/uploads/published_papers/bit04/BIT_Nah.pdf



Decrease timeouts if
hardware is too weak
for heavy loads

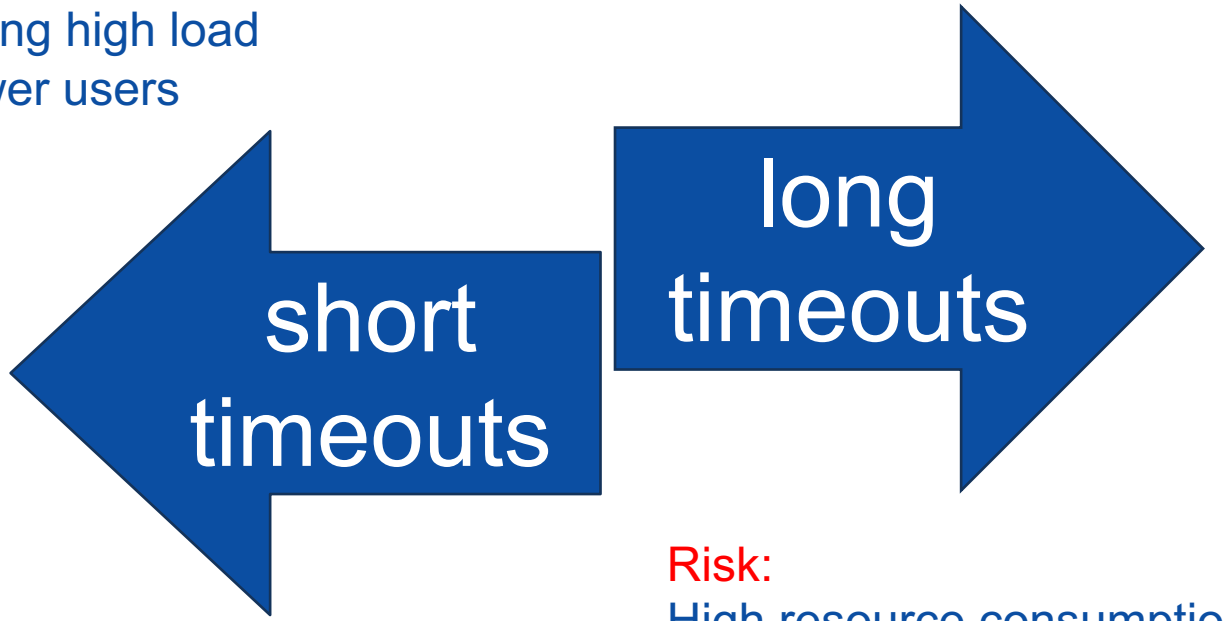
Response time depends on input (number of contracts, etc.)



Avoid low timeouts so power users can still use the system

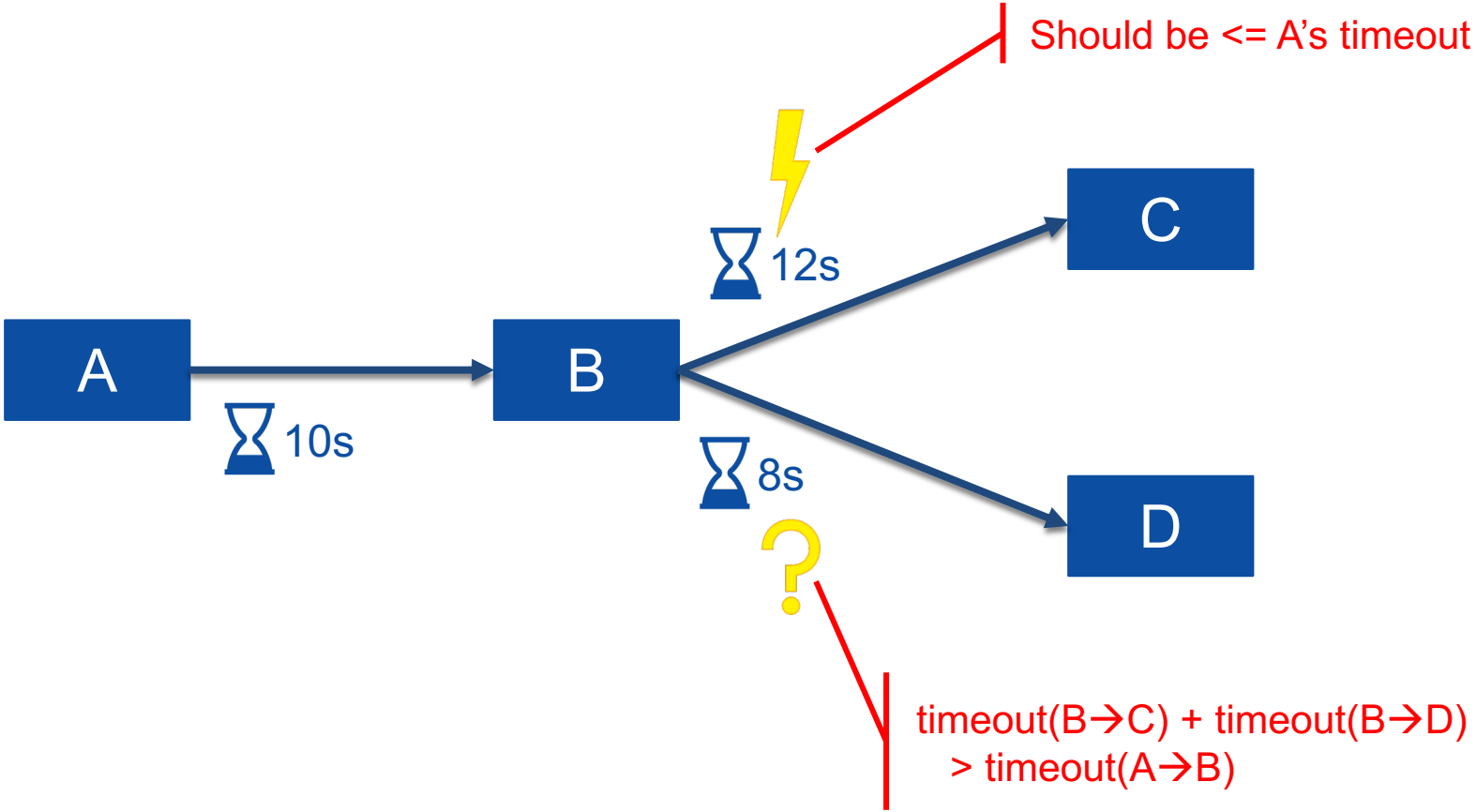
What Happens if Timeouts are too Long or too Short?

Risk:
Unusable during high load
and for power users

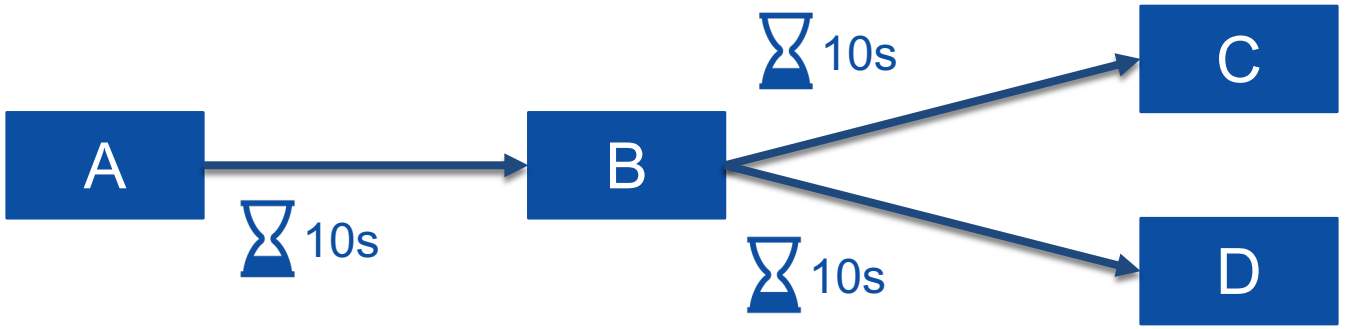


Risk:
High resource consumption
during high load (e.g. blocked threads)

Timeouts

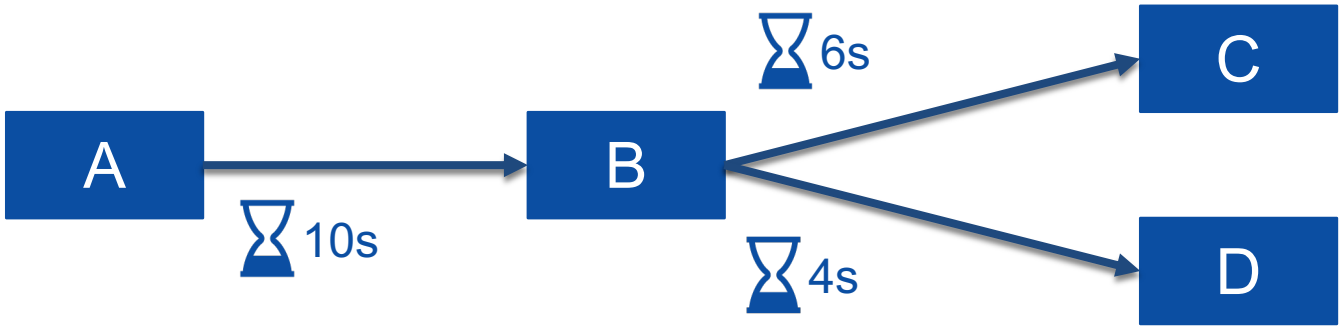

Should be <= A's timeout
12s
C
A
B
10s
8s
D
timeout(B→C) + timeout(B→D)
> timeout(A→B)

Possible Solutions



Simple

Wastes resources if C is slow

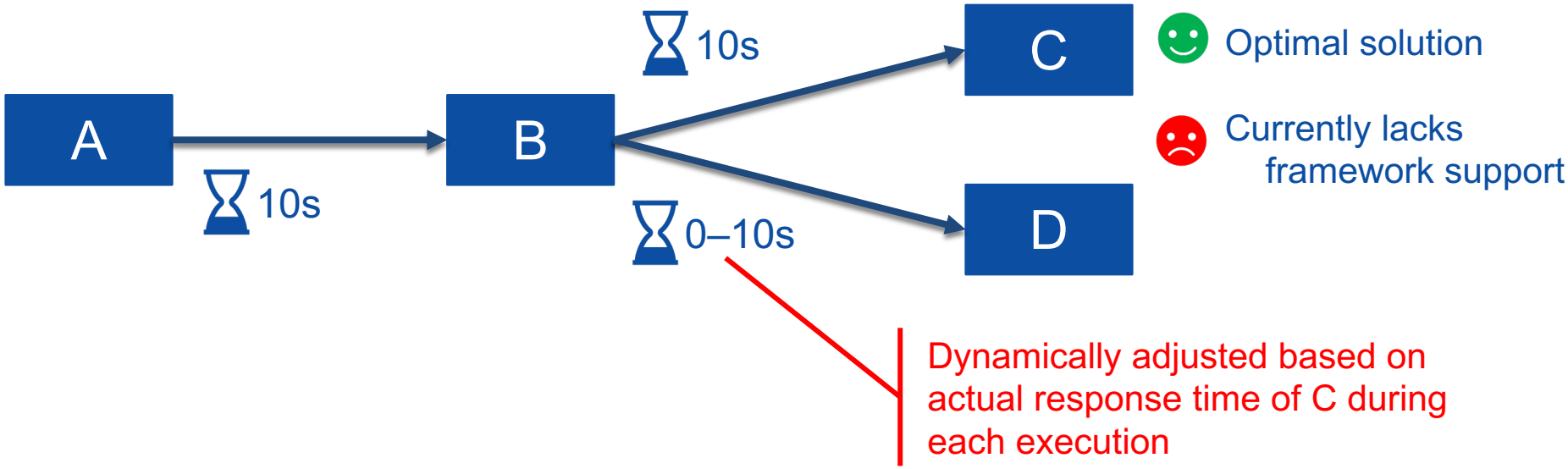


Avoids timeout at A

Difficult/impossible to find appropriate values

Needs adjustment for every new call

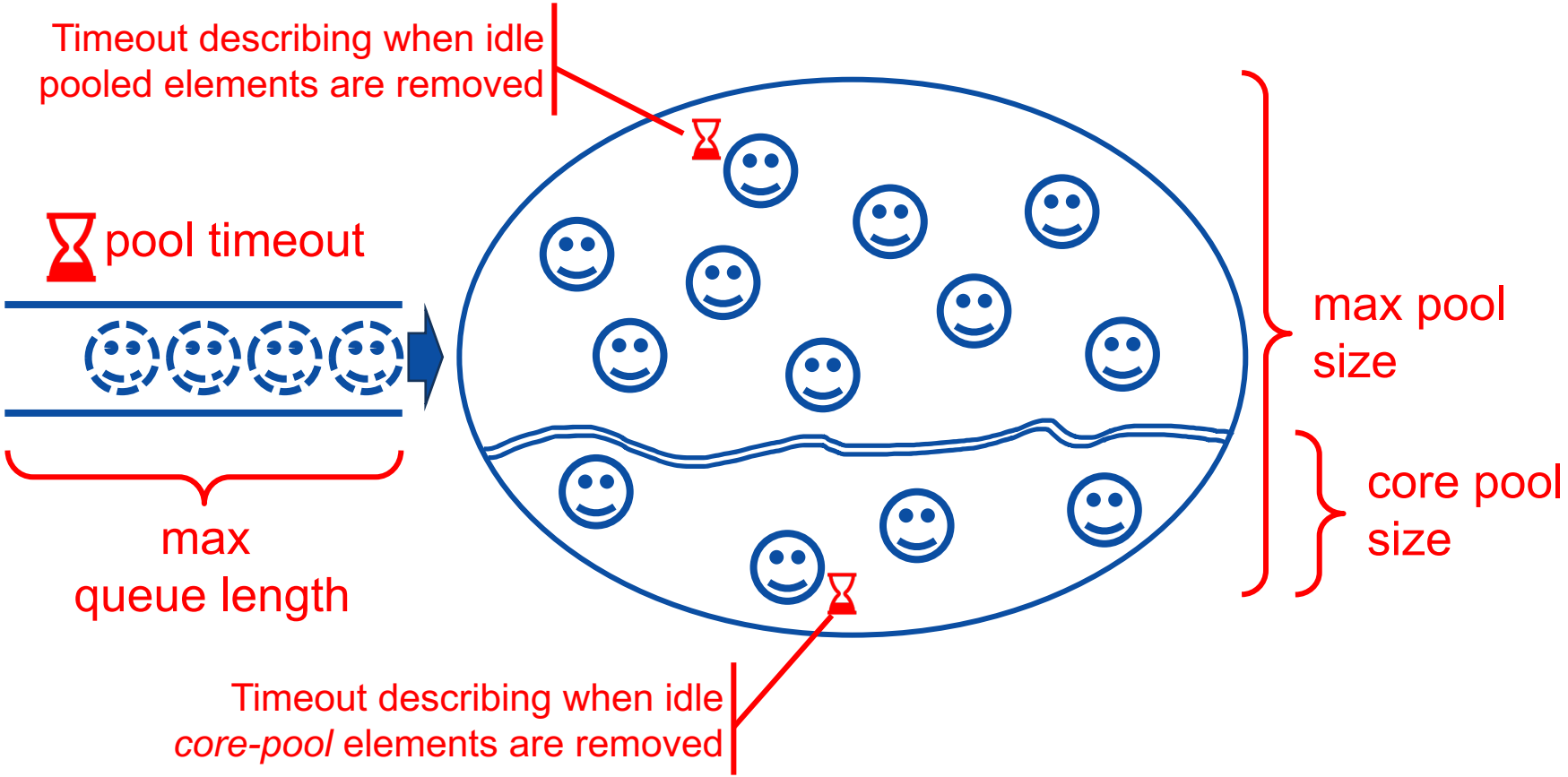
Deadline Propagation



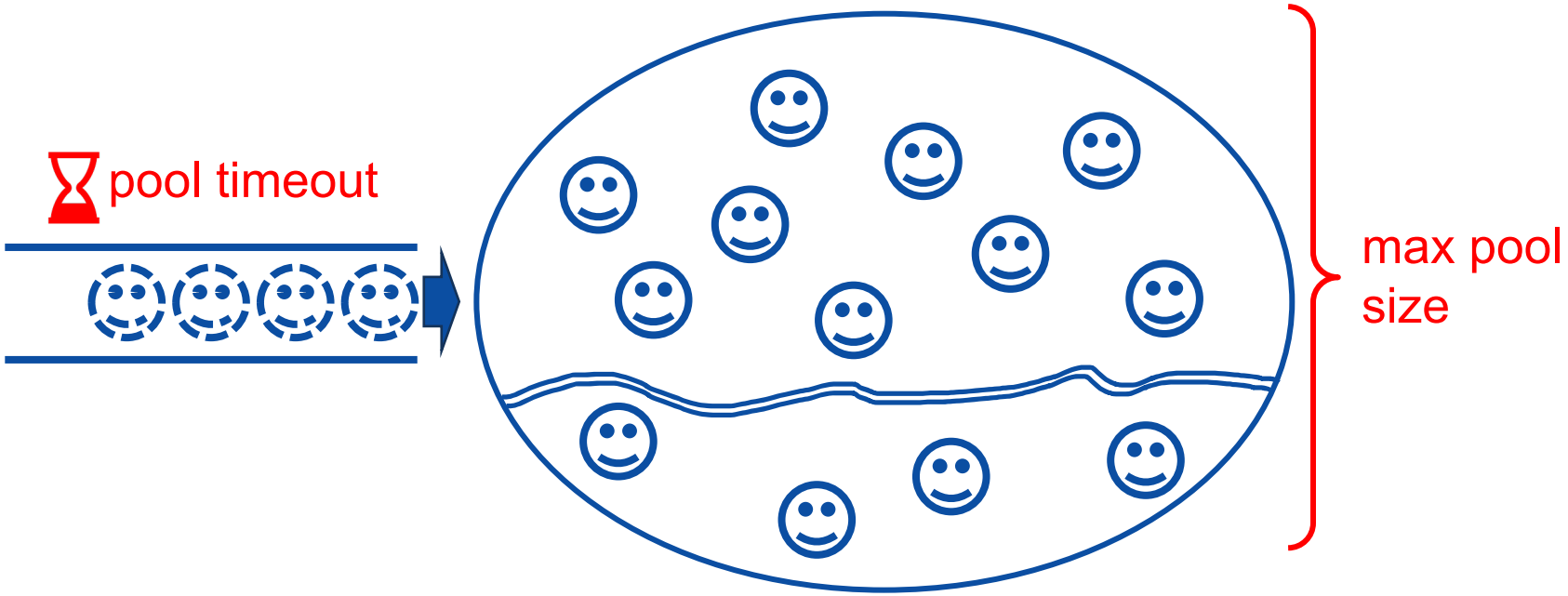
Kinds of Timeouts

	Socket Timeout	Connect Timeout	Pool Timeout (e.g. Connection Request Timeout)	Hard Timeout
How long do I wait for …	… the first byte of the answer?	… the TCP connection to establish?	… getting a pooled resource (e.g. a pooled connection)?	… the last byte of the answer?
Reason for timeout	• Other system is slow • Power User	• Firewall blocks • Other system completely overloaded • Slow network	• My system is overloaded	• Other system is slow • Power User • Long answer
Sensible values	10—20s	Much shorter	Much shorter	10—20s

Pooling


Timeout describing when idle pooled elements are removed
pool timeout
max queue length
max pool size
core pool size
Timeout describing when idle *core-pool* elements are removed

Pooling – the most important values



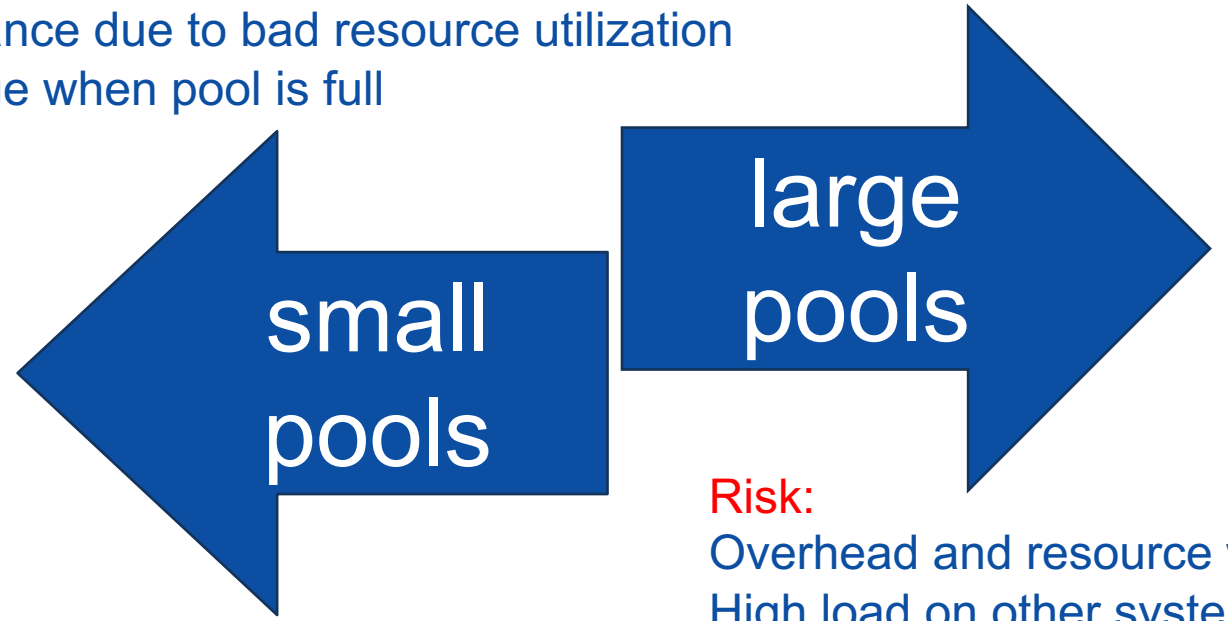
The Purpose of Pools:

Resource Acquisition

May be Expensive

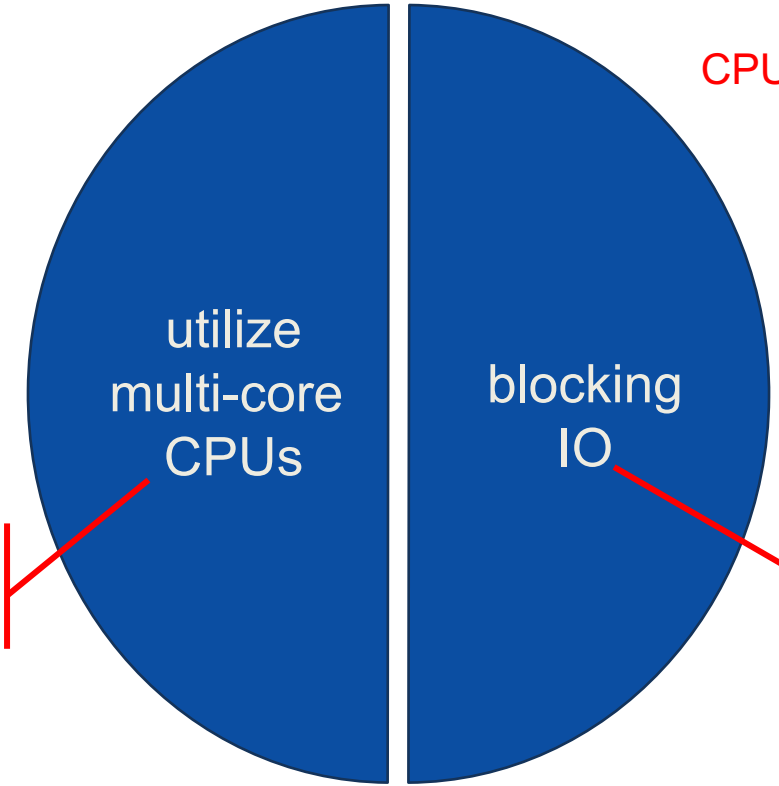
What Happens if Pools are too Small or too Large?

Risk:
Low performance due to bad resource utilization
Service outage when pool is full

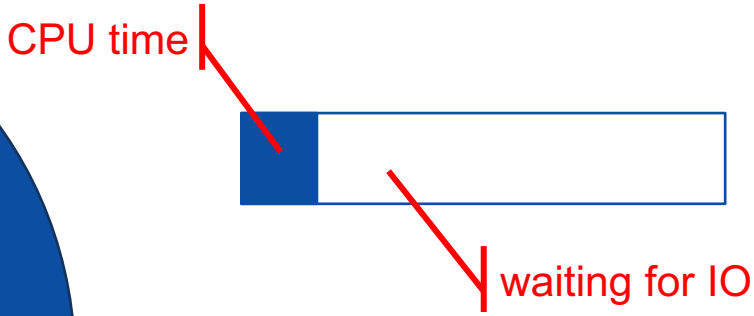


Risk:
Overhead and resource waste
High load on other systems
Requests queue up
Slow recovery from high load

The Purpose of Threads

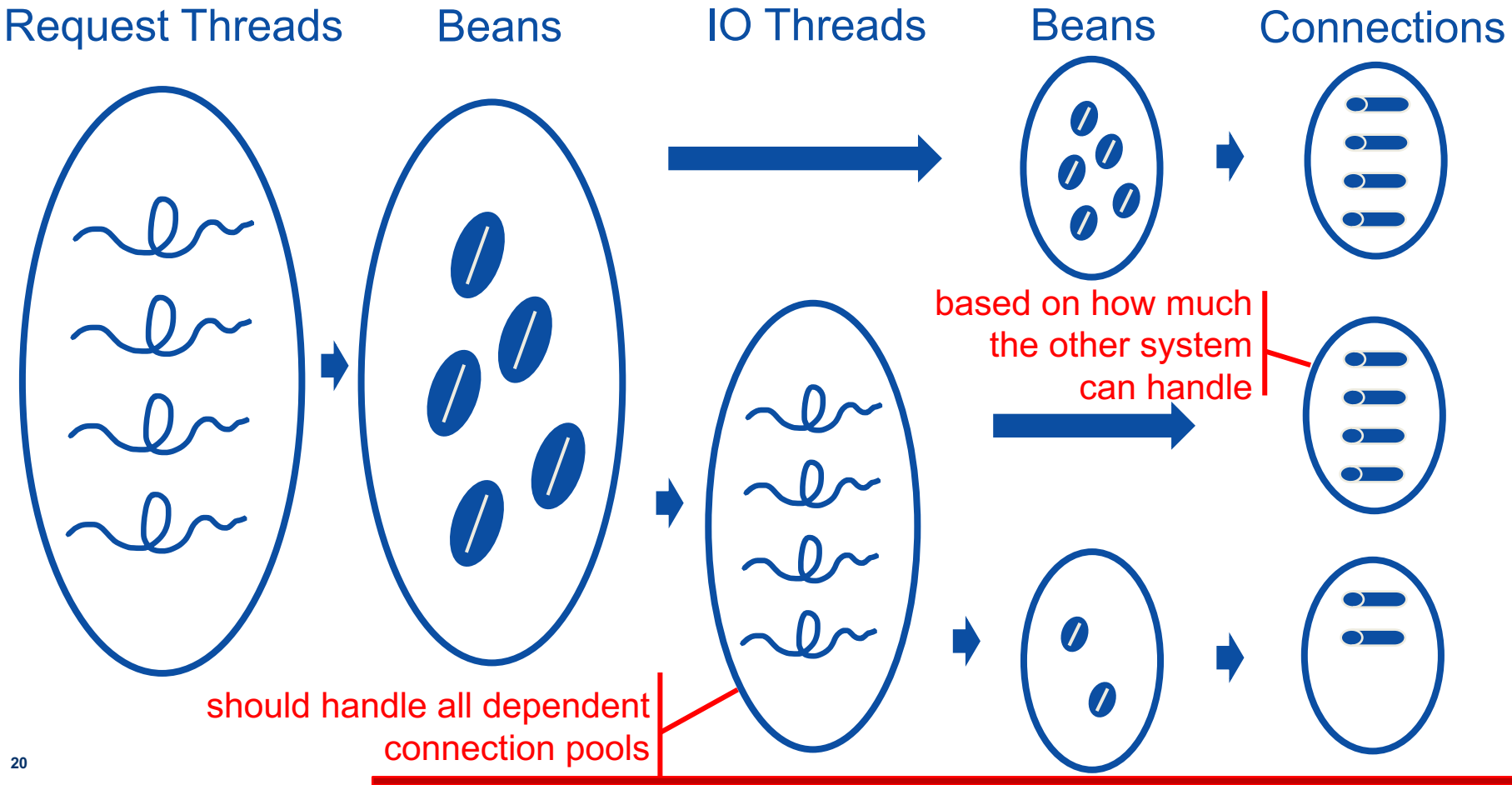


pool-size based on
number of CPU cores

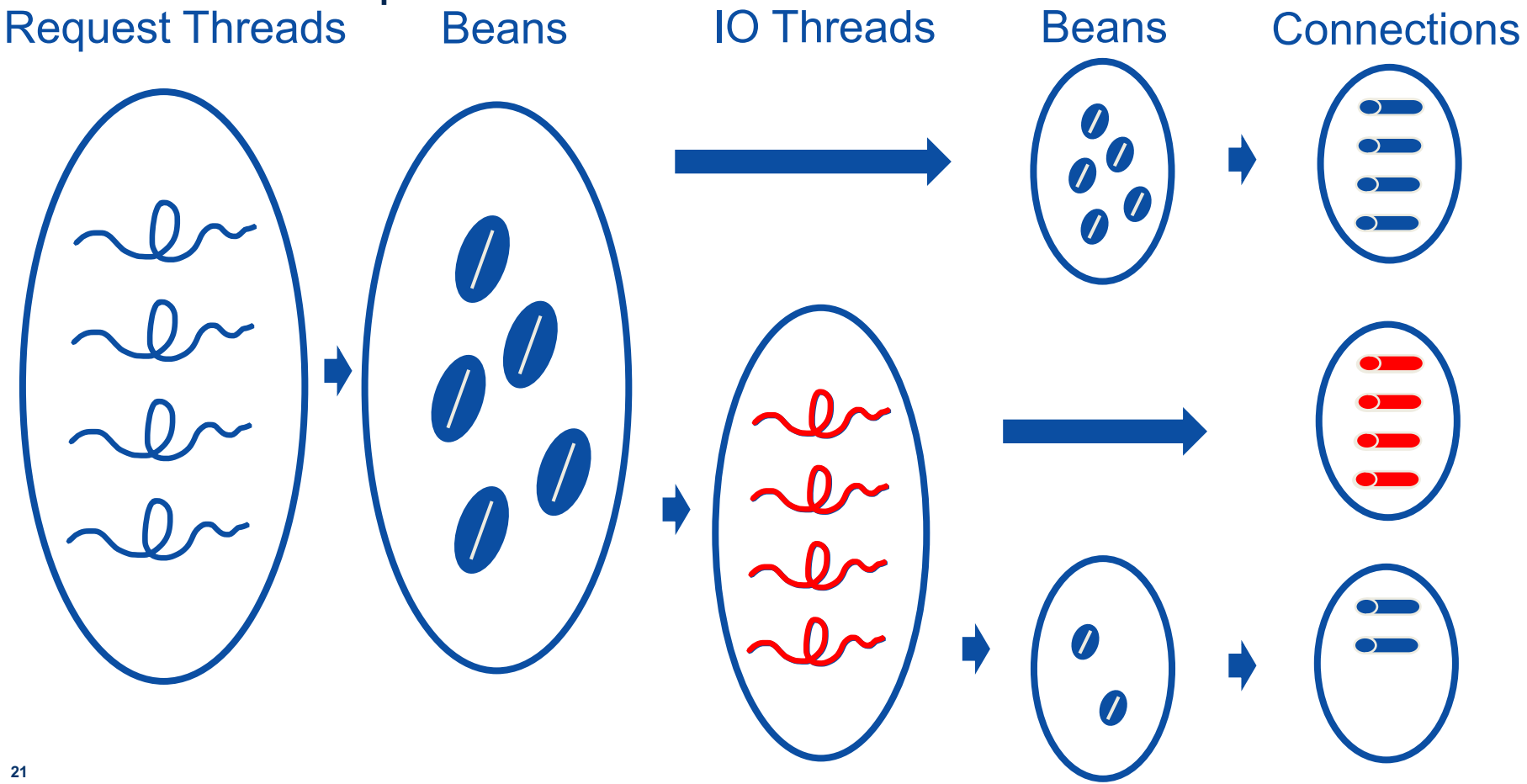


If CPU time is 5% then
pool-size = 20 * numberOfCores

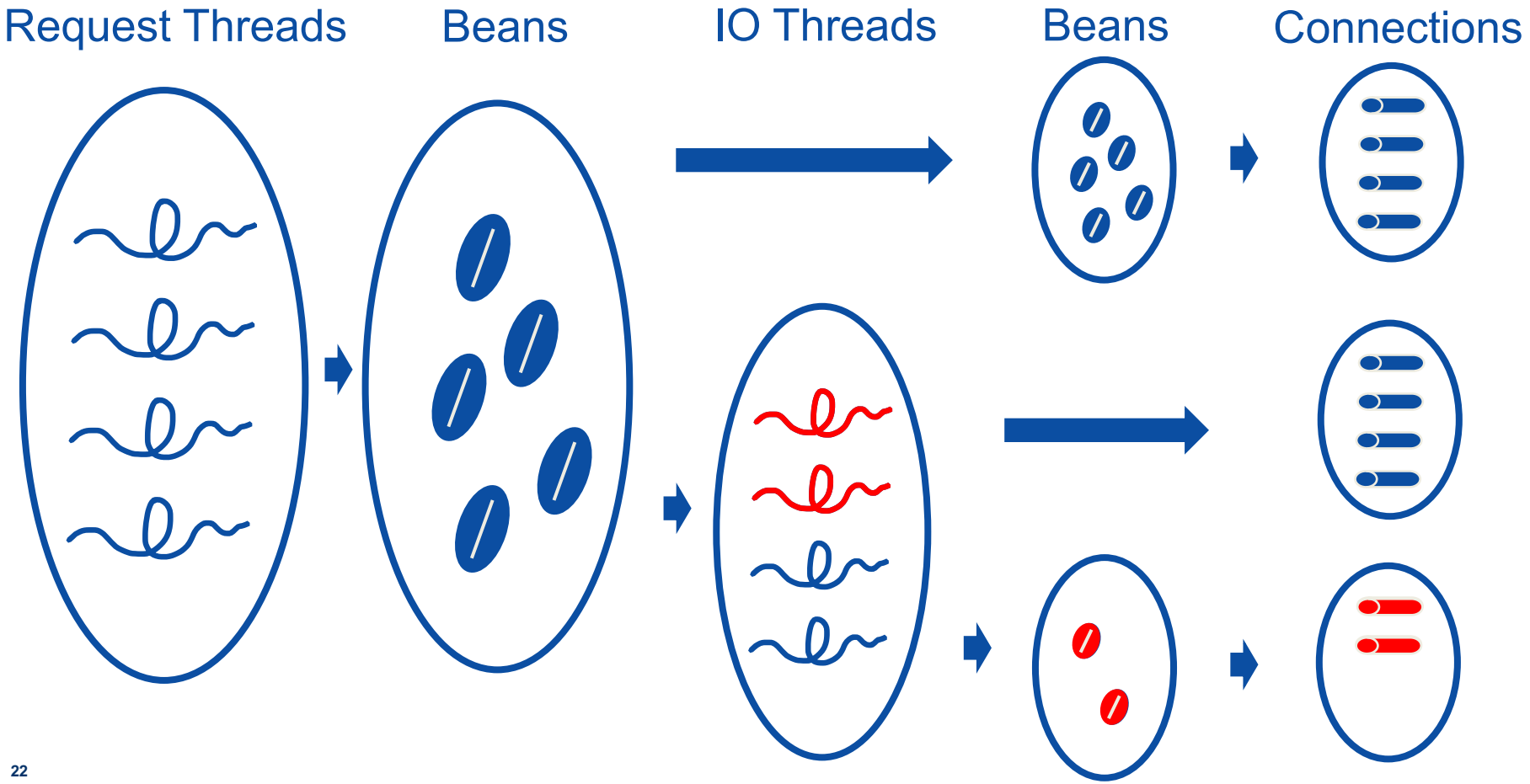
The Whole Picture



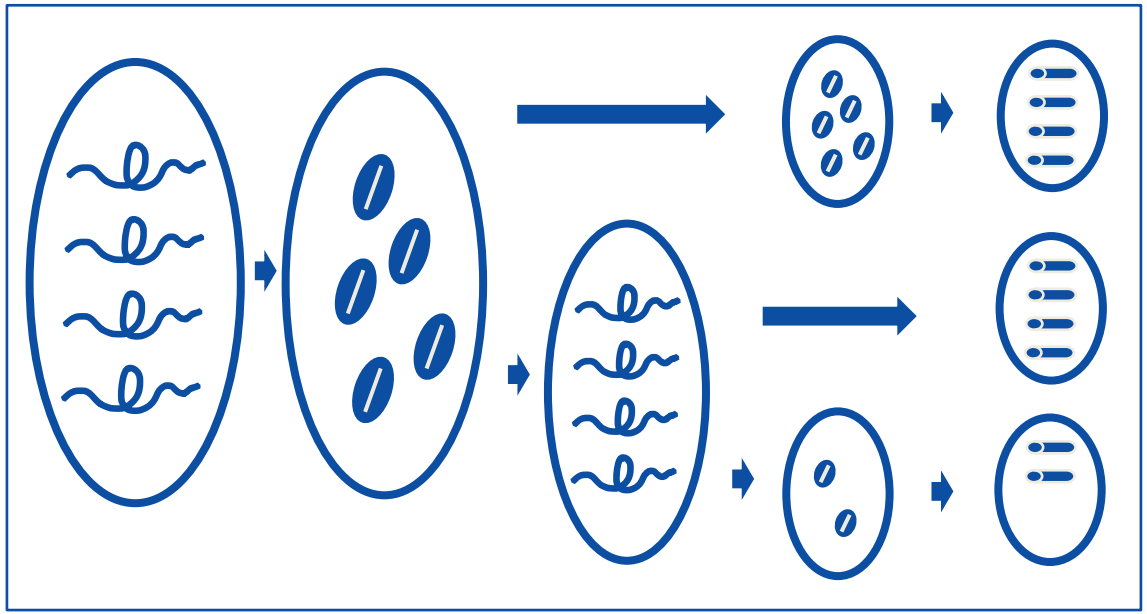
The Thread Pool Fills up

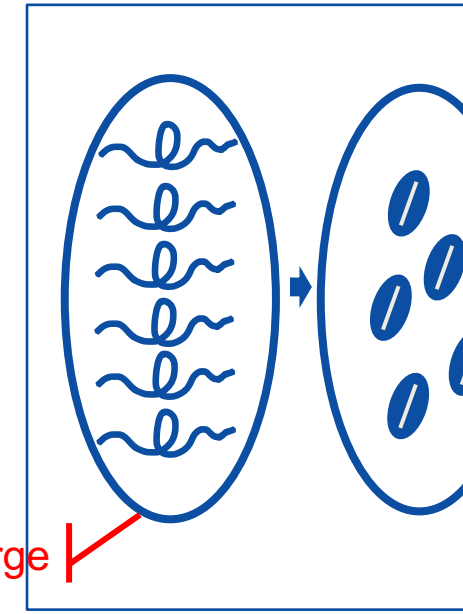


Bulk Head with Beans

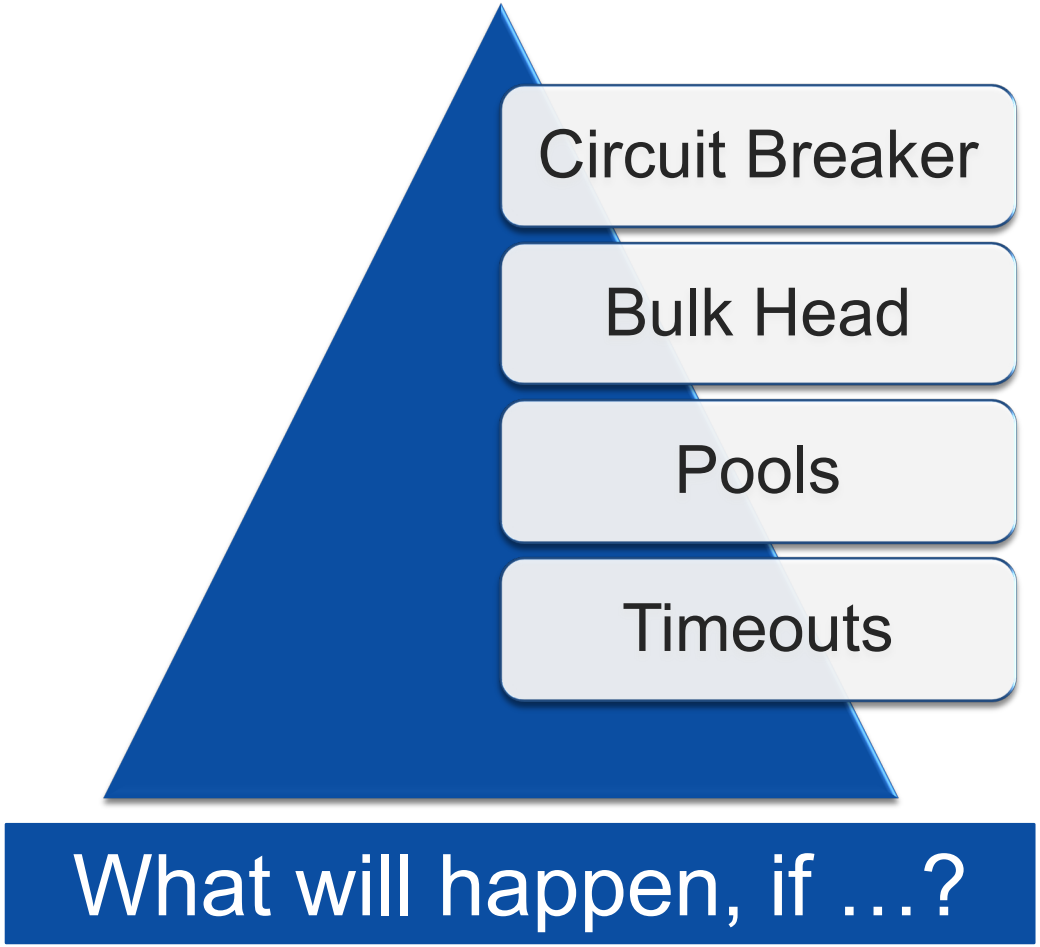


Circuit Breaker





Order of Priority



Thank You!

Questions?

This presentation in written form

http://www.christian-rehn.de/2019/10/03/die-praesidentenmacherin-und-der-timeout/