Pipelining in Engineering

ENGR 102

Keith M. Chugg
Fall 2015

USC University of Southern California

Overview

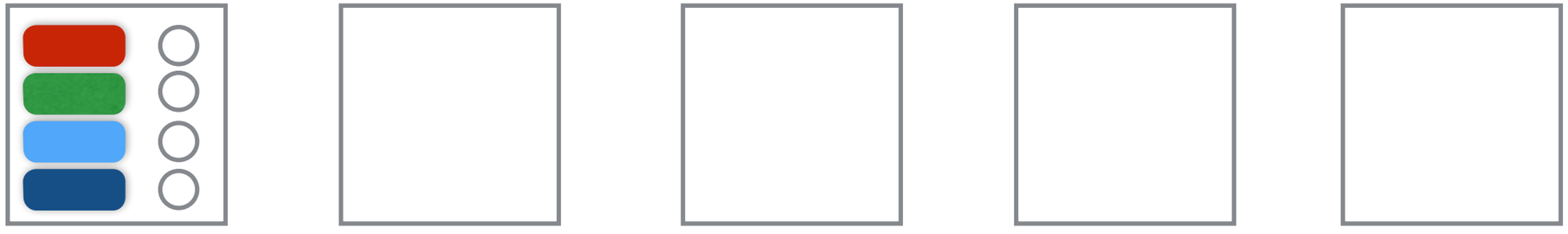
- Explore the concept of pipelining through an example
  - Group reading
  - Some reality checks
    - Refining our models
    - Limitations of pipeling
- Characterizing general pipelining approaches

Group Reading

- 4-page article
- 5 students in the class
- Each student can read 1 page/minute
- *How long does it take for everybody to ready the article?*

Group Reading - Direct Approach

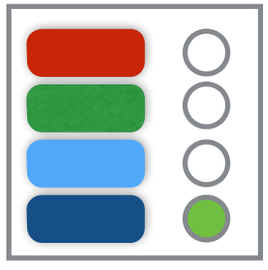
T = 0



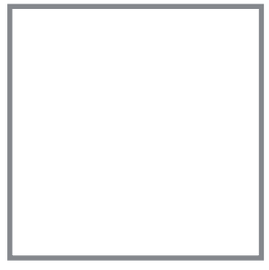
page 4
page 3
page 2
page 1

Group Reading - Direct Approach

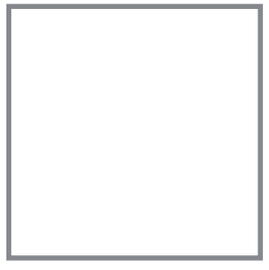
T = 1



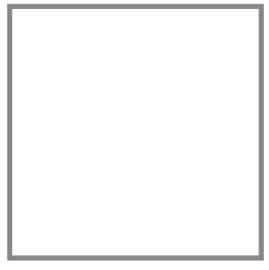
student 1



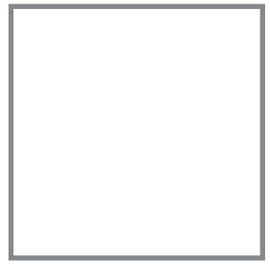
student 2



student 3



student 4

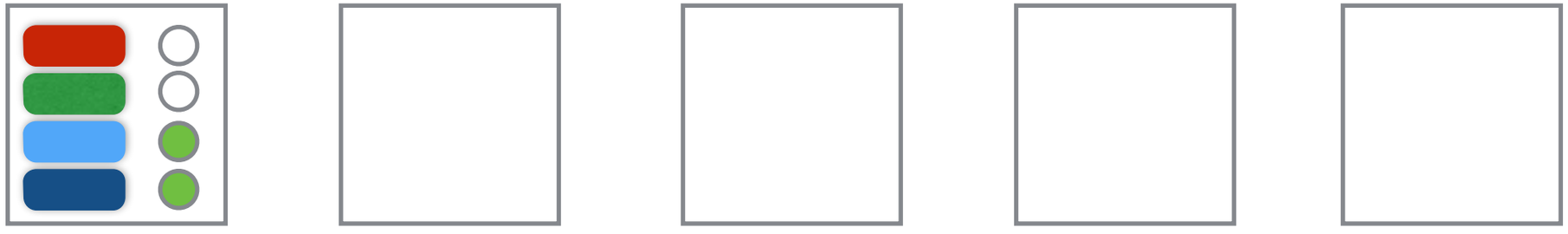


student 5

page 4
page 3
page 2
page 1

Group Reading - Direct Approach

$T = 2$

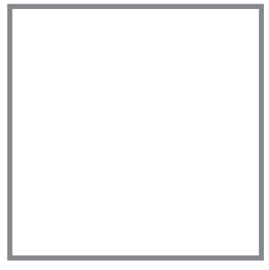


page 4
page 3
page 2
page 1

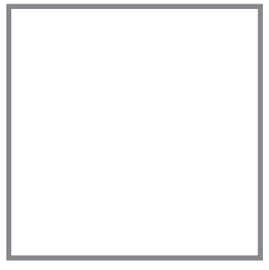
Group Reading - Direct Approach

T = 3

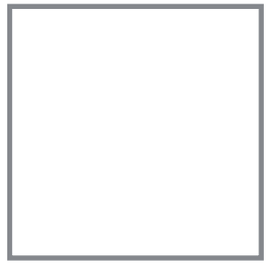
student 1



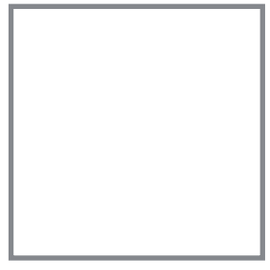
student 2



student 3



student 4

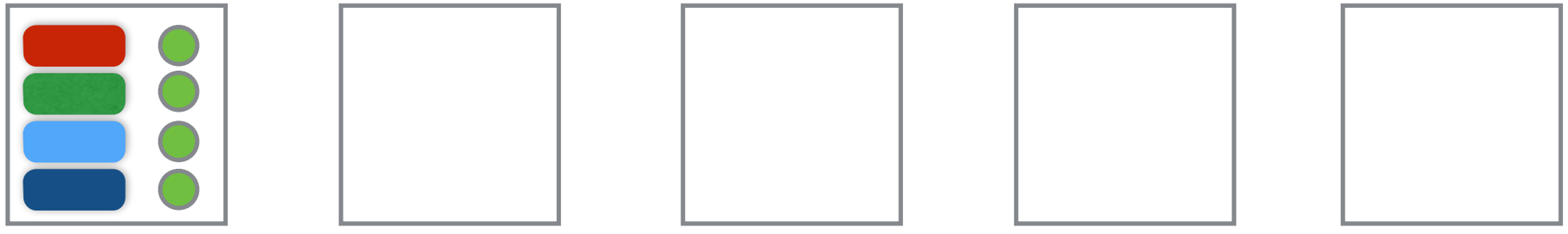


student 5

page 4
page 3
page 2
page 1

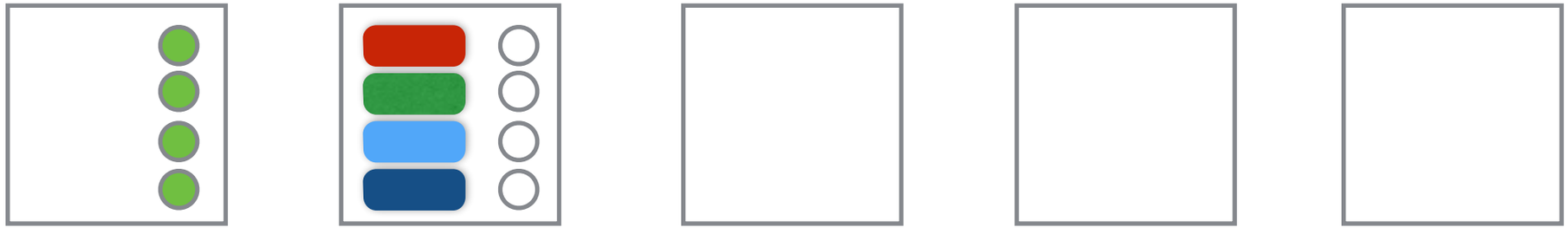
Group Reading - Direct Approach

T = 4



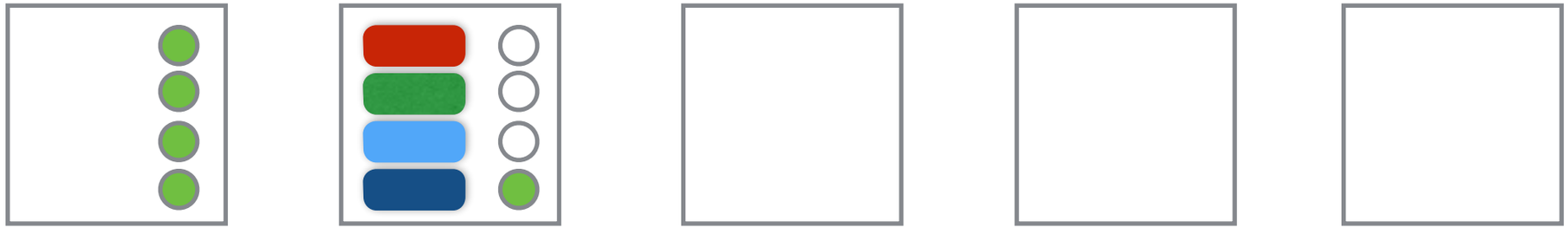
Group Reading - Direct Approach

T = 5



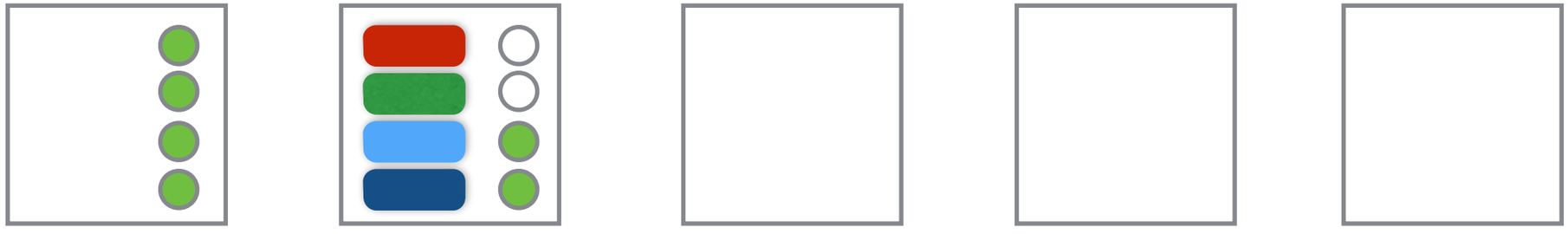
Group Reading - Direct Approach

T = 5



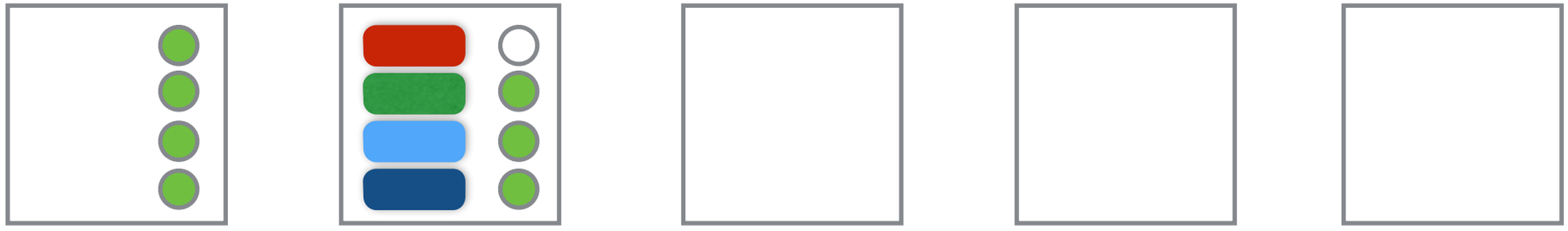
Group Reading - Direct Approach

T = 6



Group Reading - Direct Approach

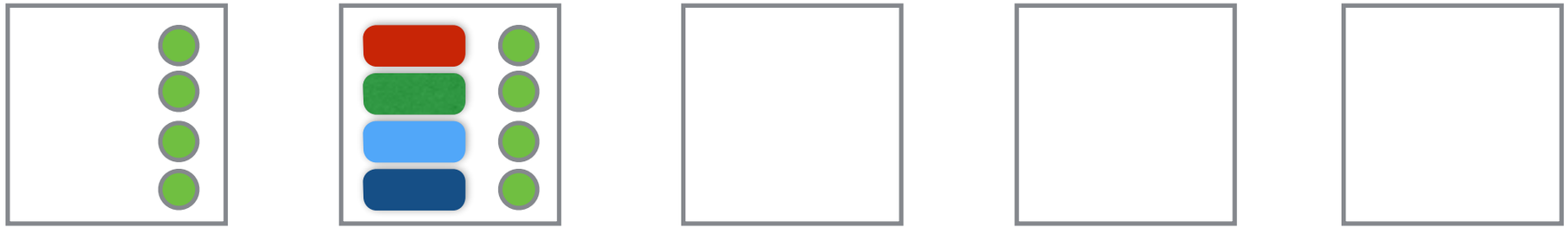
T = 7



page 4
page 3
page 2
page 1

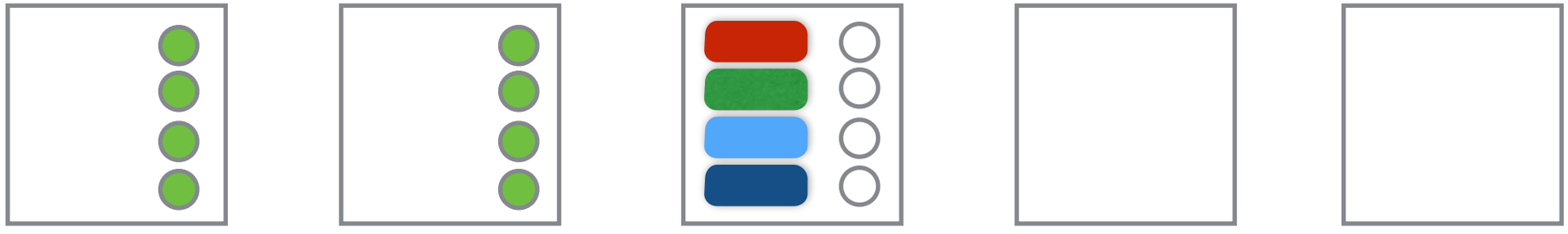
Group Reading - Direct Approach

T = 8



Group Reading - Direct Approach

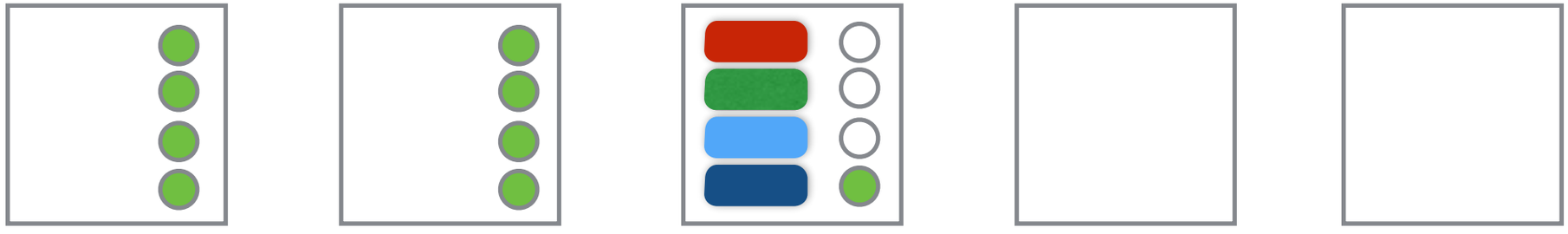
T = 9



page 4
page 3
page 2
page 1

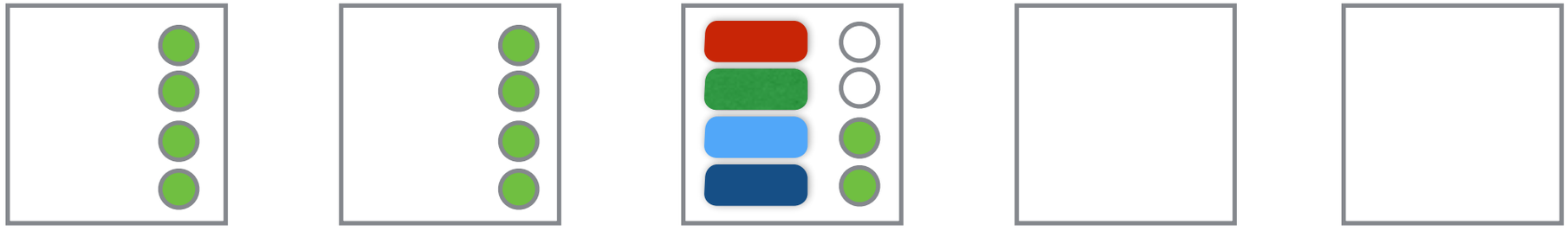
Group Reading - Direct Approach

T = 9



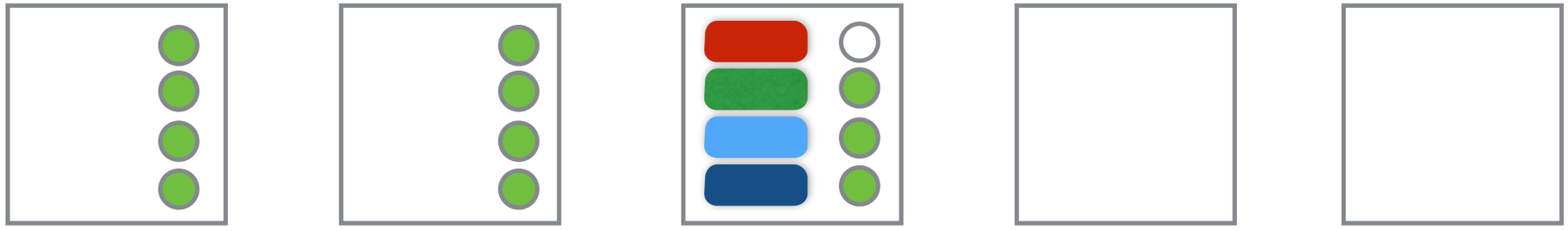
Group Reading - Direct Approach

T = 10



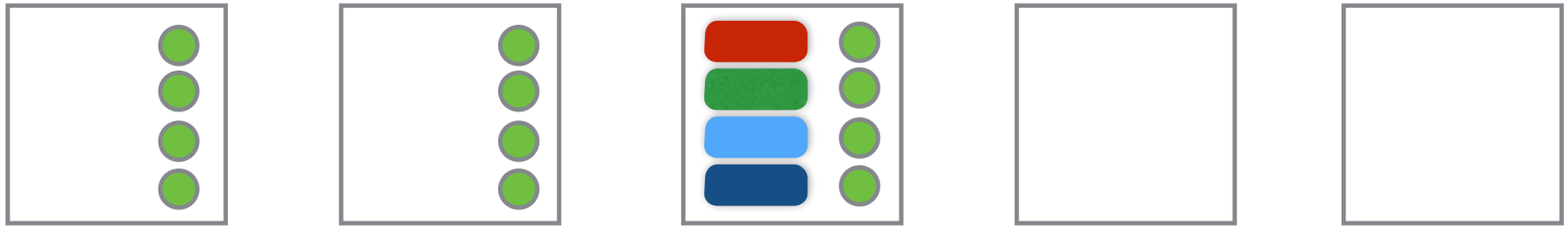
Group Reading - Direct Approach

$T = 11$



Group Reading - Direct Approach

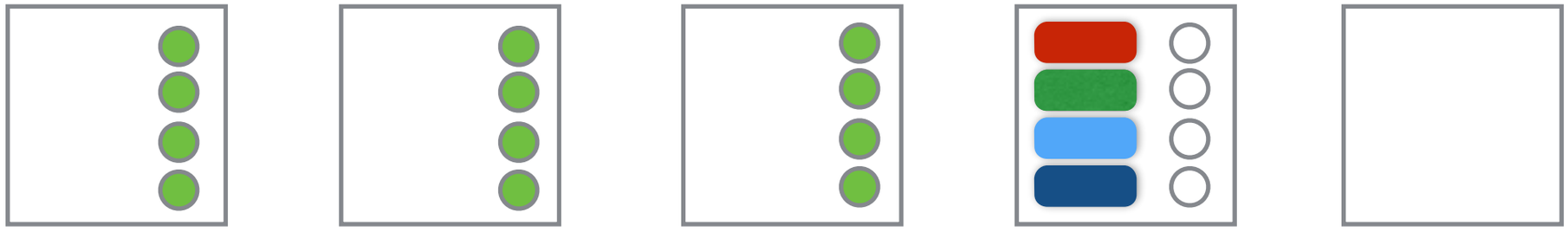
T = 12



page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 13

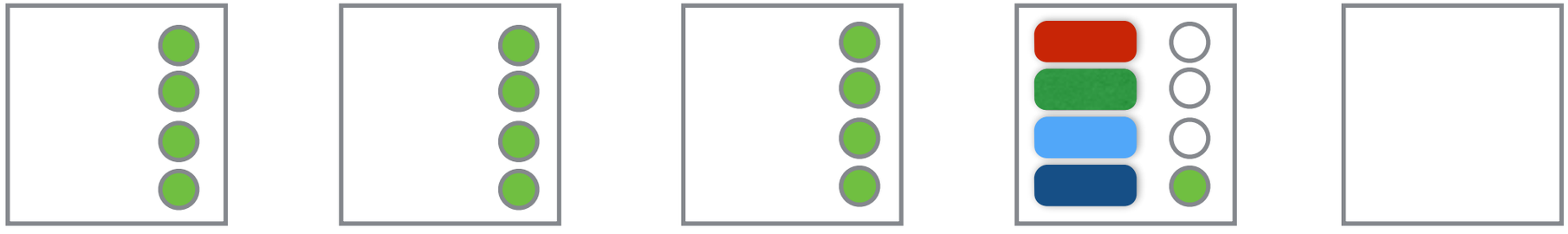


student 1 | student 2 | student 3 | student 4 | student 5

page 4
page 3
page 2
page 1

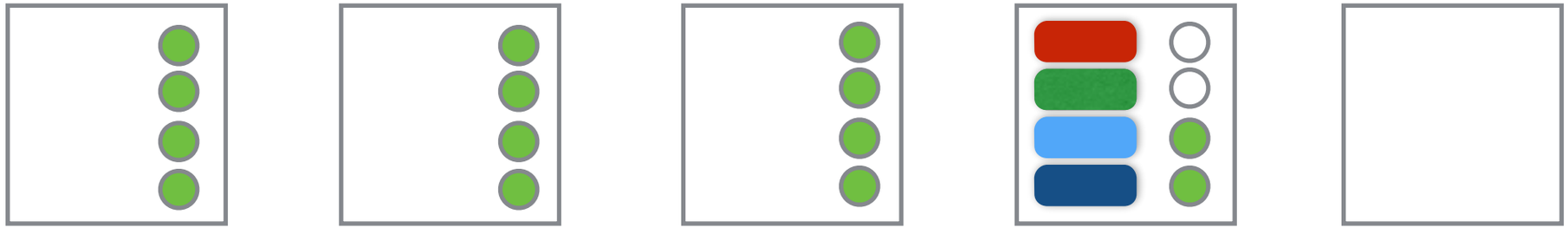
Group Reading - Direct Approach

T = 13



Group Reading - Direct Approach

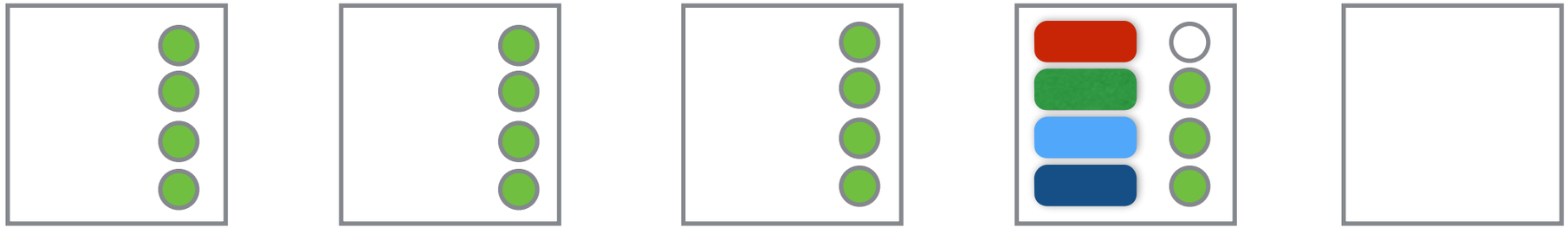
T = 14



page 4
page 3
page 2
page 1

Group Reading - Direct Approach

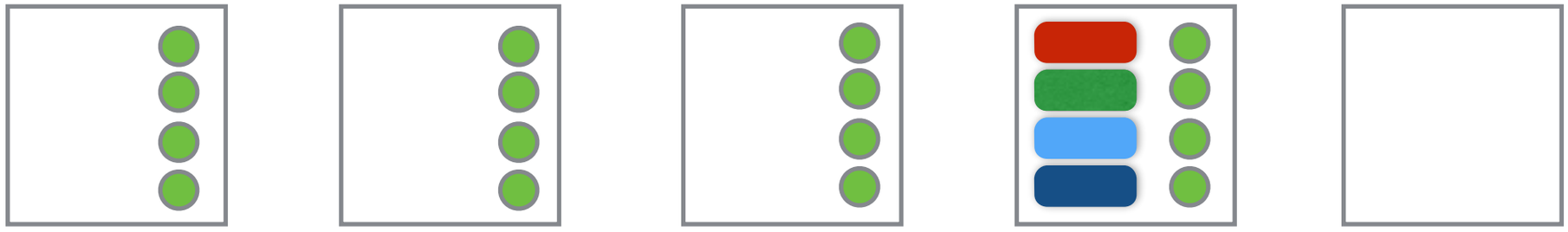
T = 15



page 4
page 3
page 2
page 1

Group Reading - Direct Approach

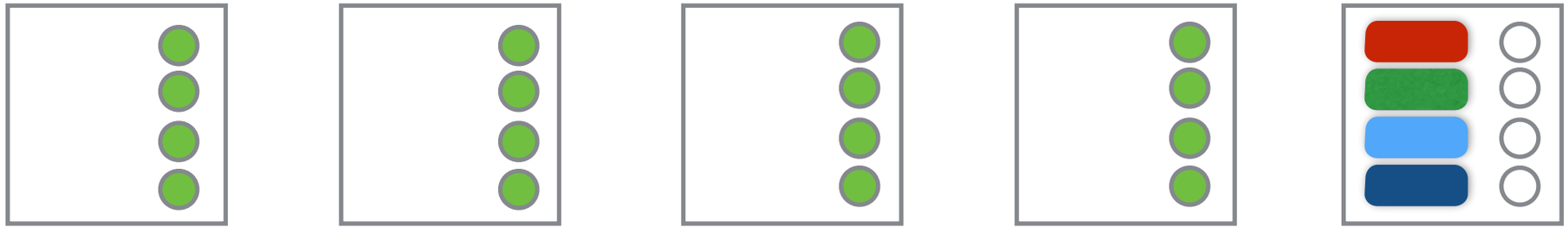
T = 16



page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 17

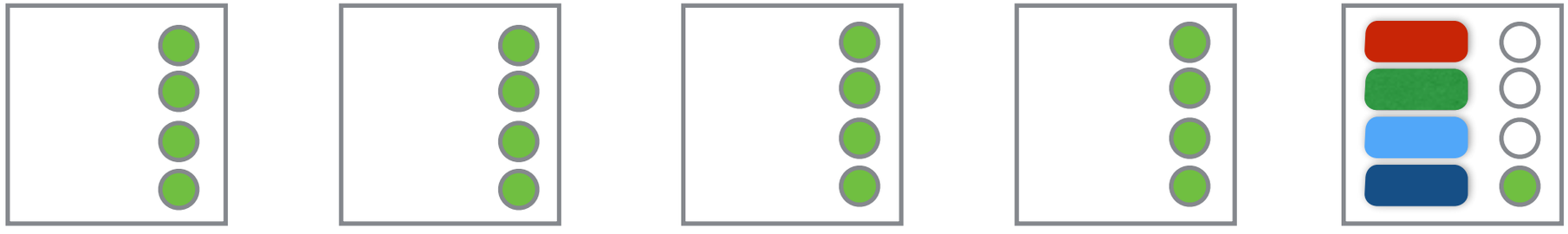


student 1 | student 2 | student 3 | student 4 | student 5

page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 17

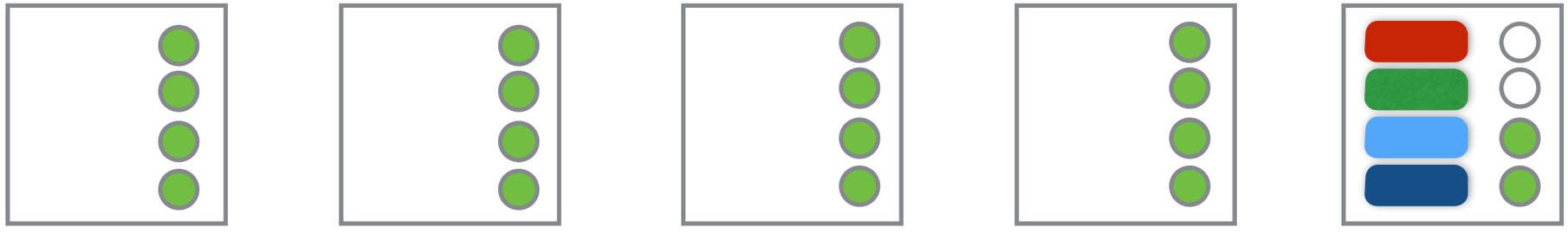


student 1 | student 2 | student 3 | student 4 | student 5

page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 18

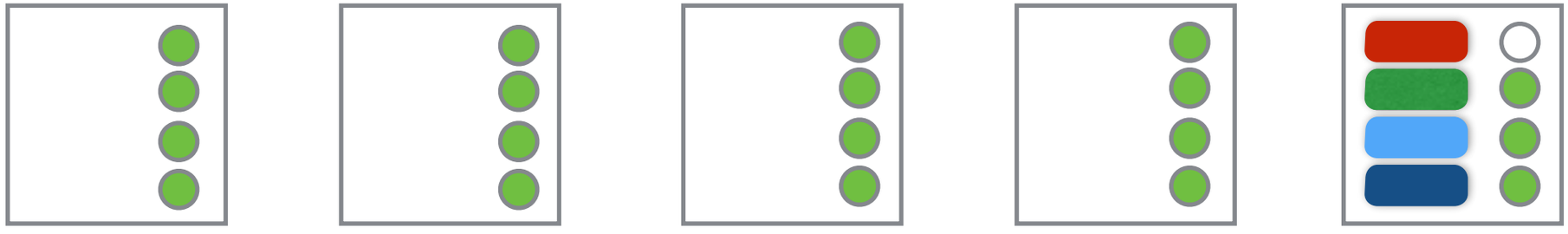


student 1 | student 2 | student 3 | student 4 | student 5

page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 19

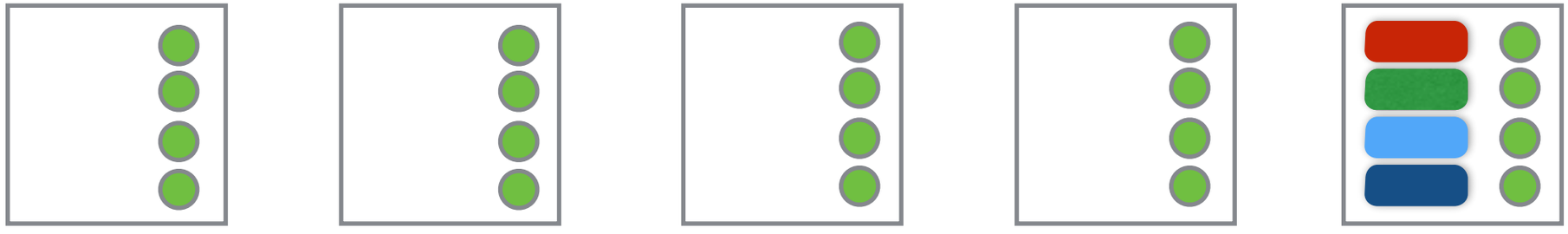


student 1 student 2 student 3 student 4 student 5

page 4
page 3
page 2
page 1

Group Reading - Direct Approach

T = 20



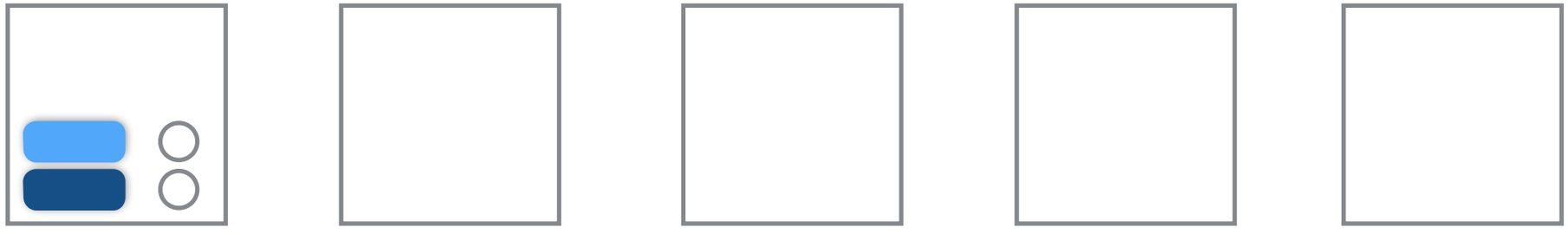
- *20 time units to complete (mins)*
- *Each student requires 4 time units to complete*
- *Time between one student and the next finishing is 4 time units (mins)*

How to speed this up?

- What is slowing us down?
- Distinct methods/concepts:
  - *Parallelism*
  - *Division of labor (job partitioning)*
  - *Pipelining*

Group Reading - Pipeline with Two Sections

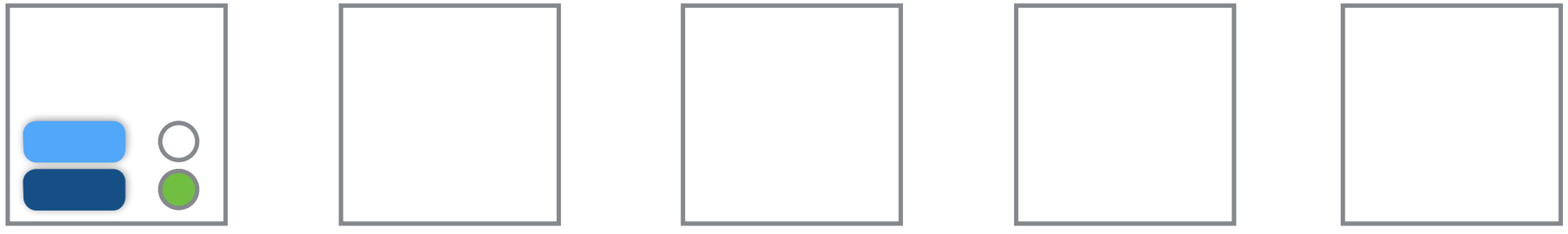
T = 0



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

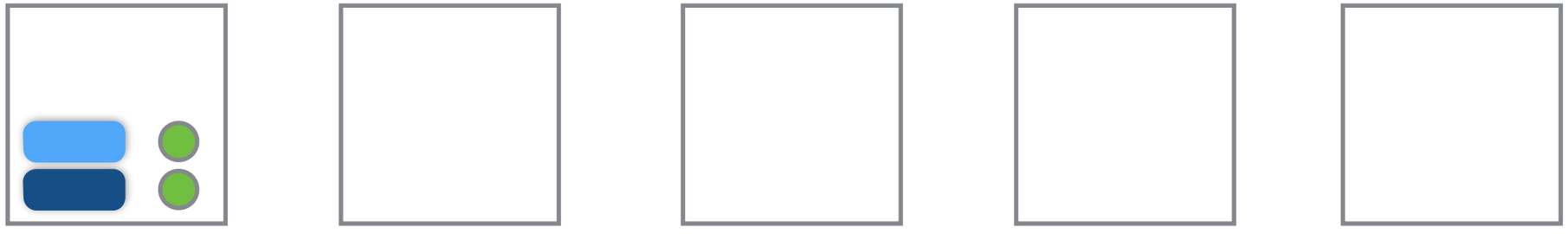
T = I



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

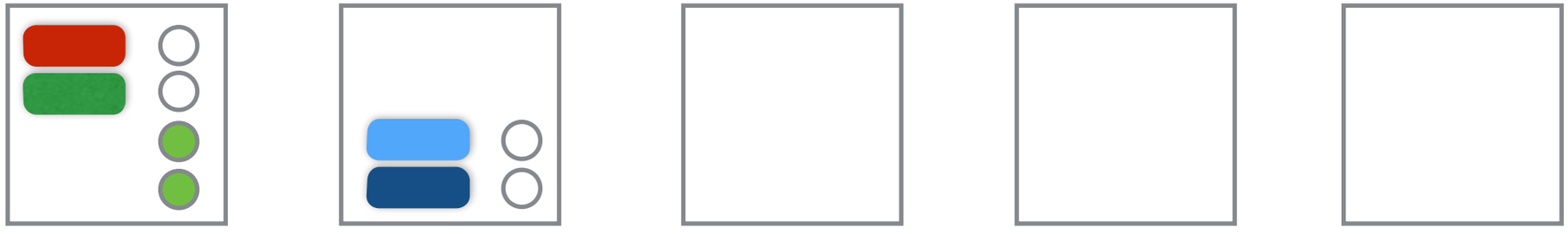
T = 2



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

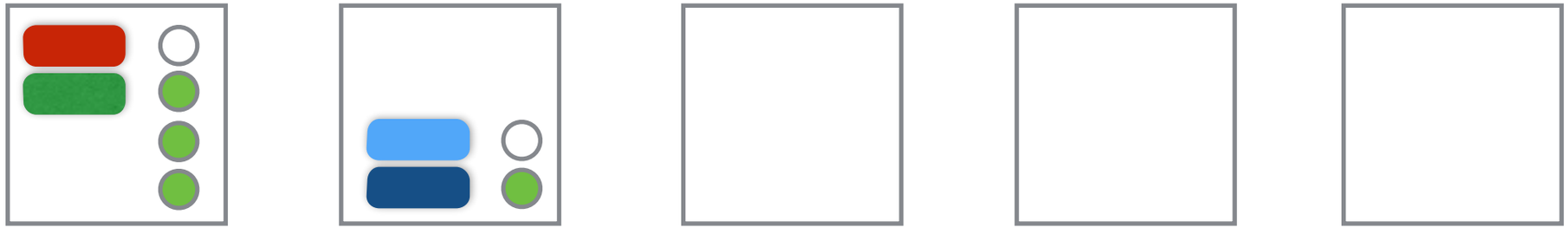
T = 3



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

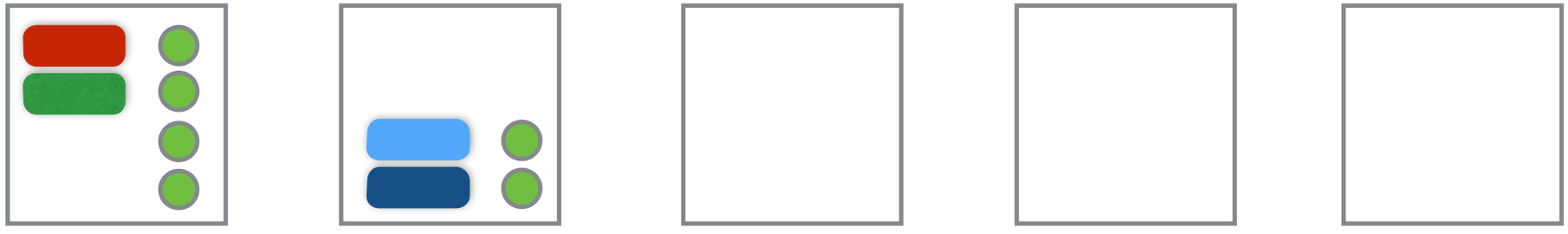
T = 3



page 4
page 3
page 2
page 1

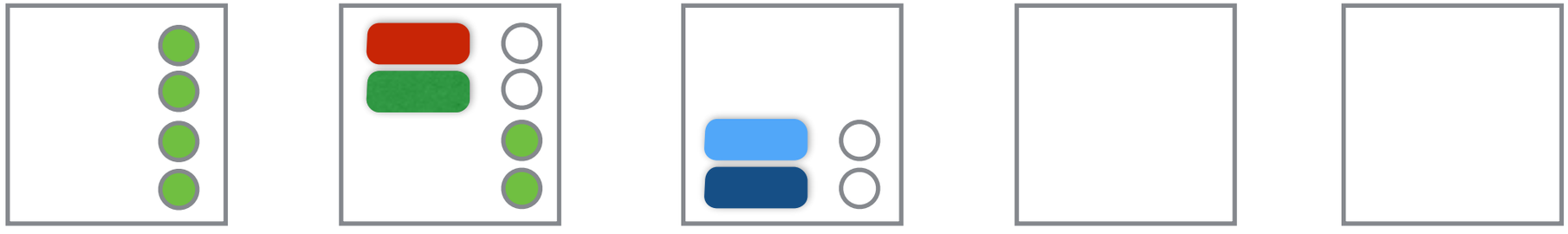
Group Reading - Pipeline with Two Sections

T = 4



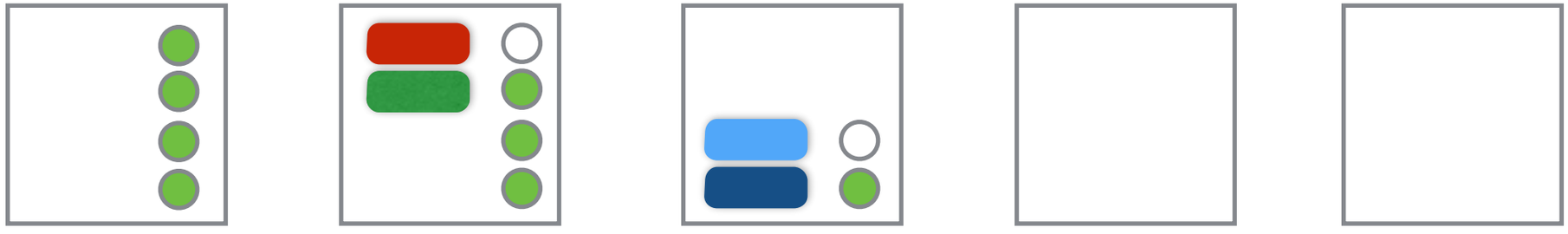
Group Reading - Pipeline with Two Sections

T = 5



Group Reading - Pipeline with Two Sections

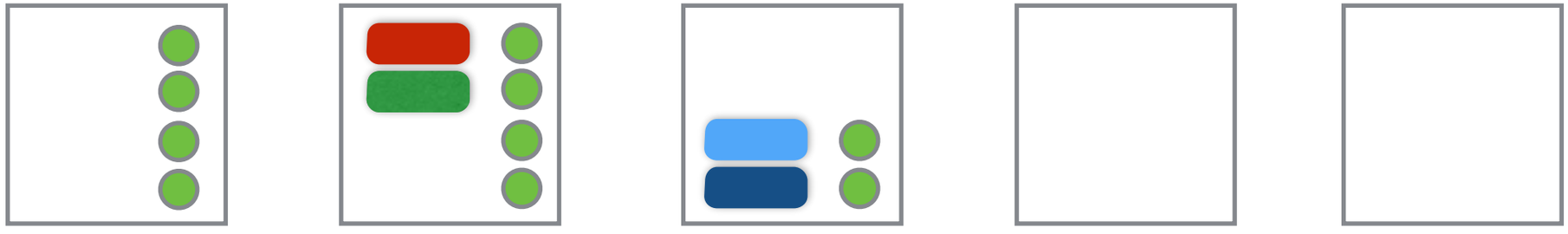
T = 5



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

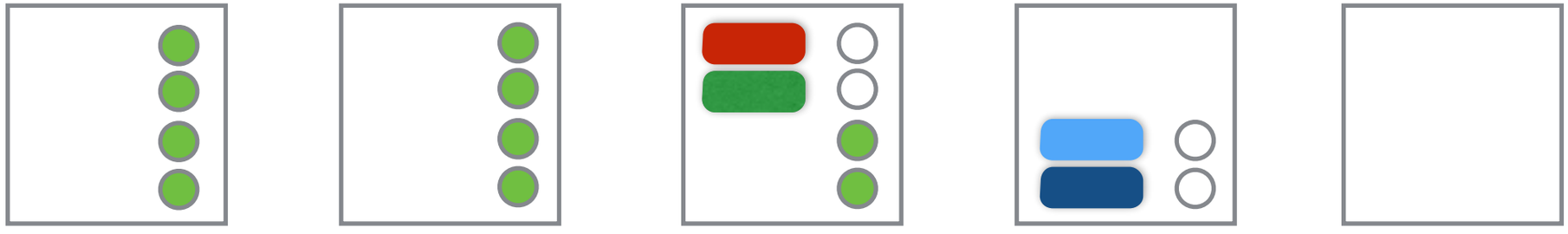
T = 6



page 4
page 3
page 2
page 1

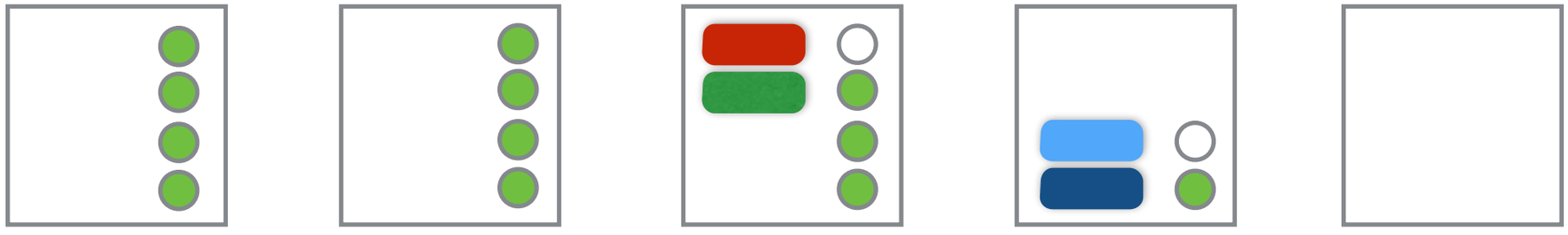
Group Reading - Pipeline with Two Sections

T = 7



Group Reading - Pipeline with Two Sections

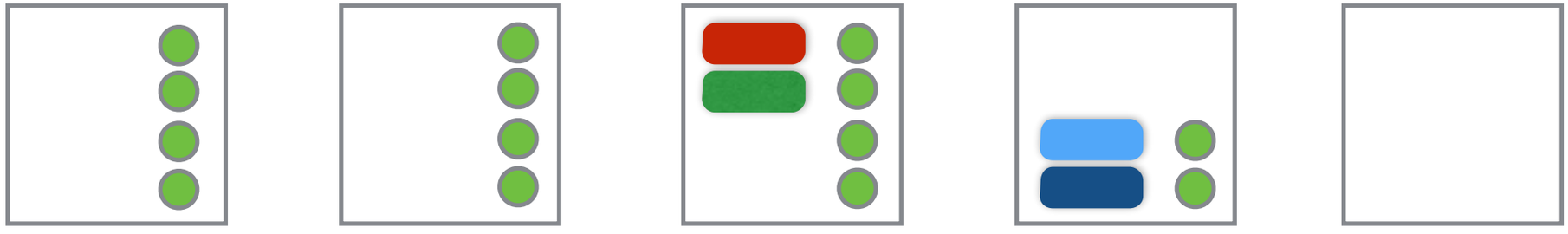
T = 7



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

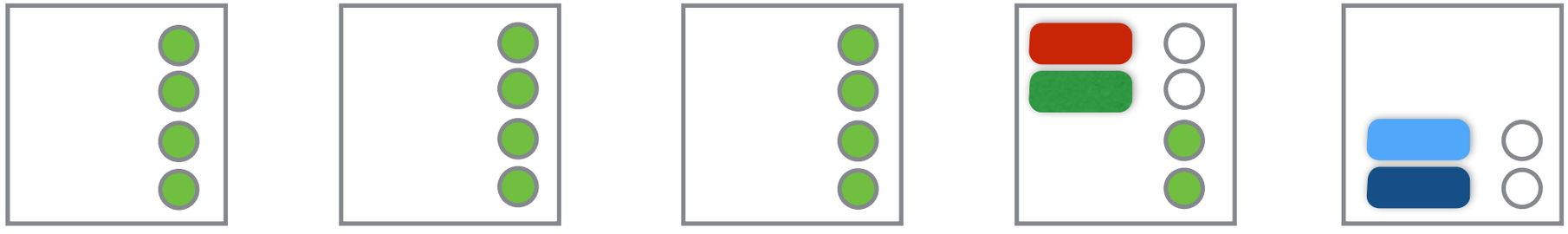
T = 8



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

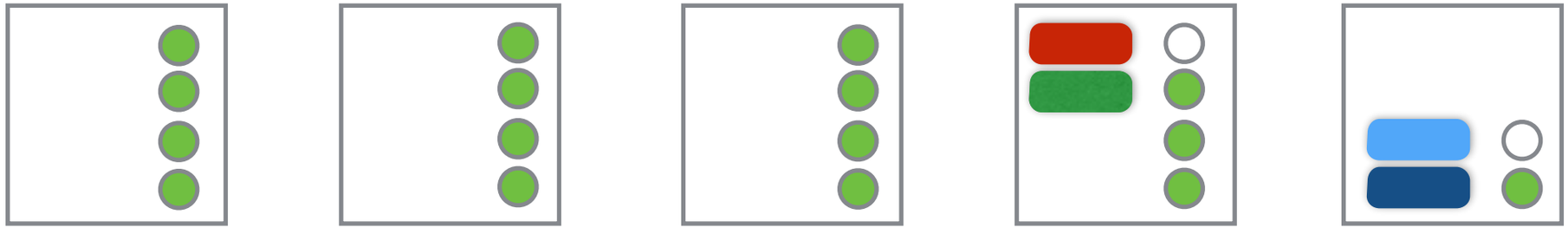
T = 9



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

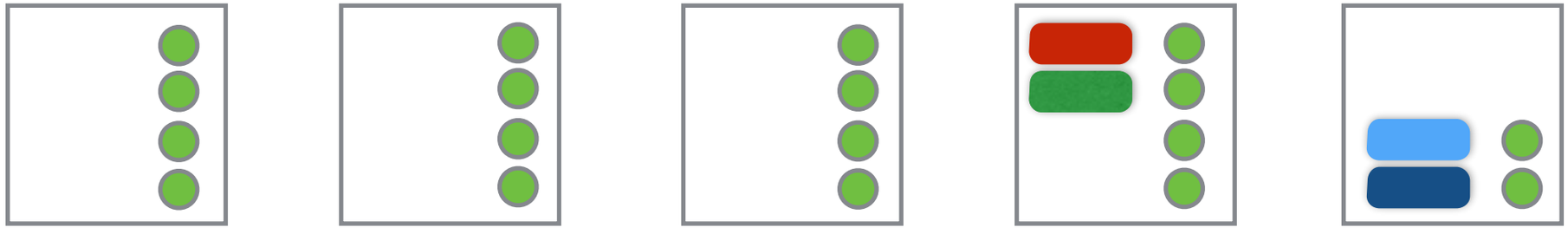
T = 9



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

T = 10

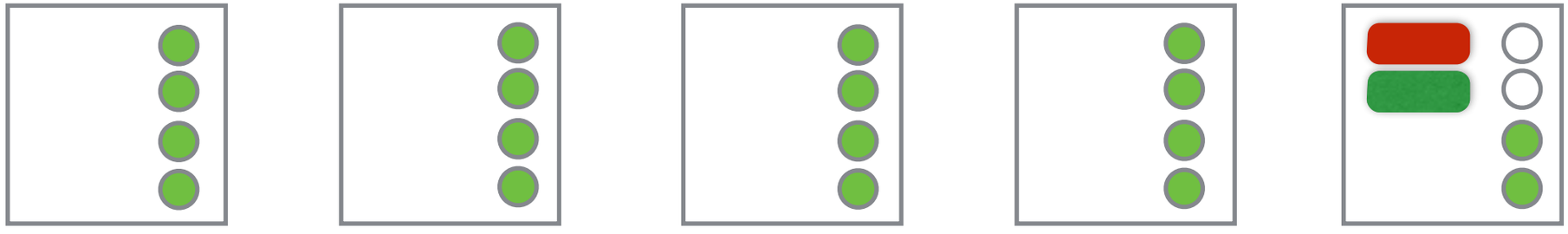


student 1 student 2 student 3 student 4 student 5

page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

T = 11

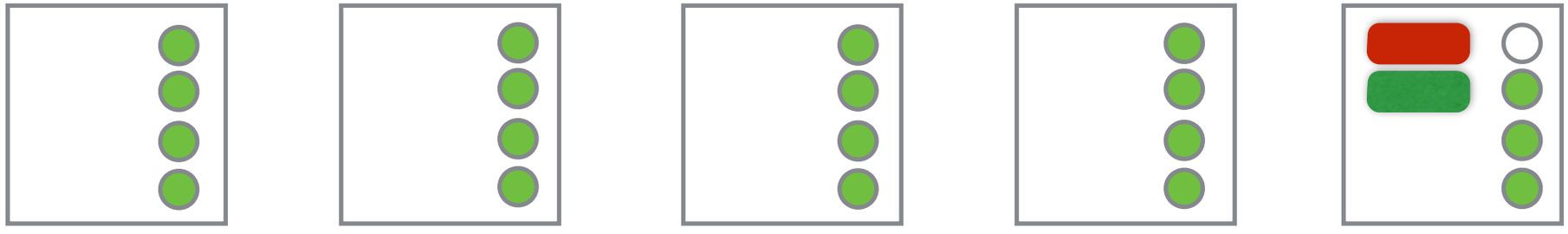


student 1 student 2 student 3 student 4 student 5

page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

$T = 11$

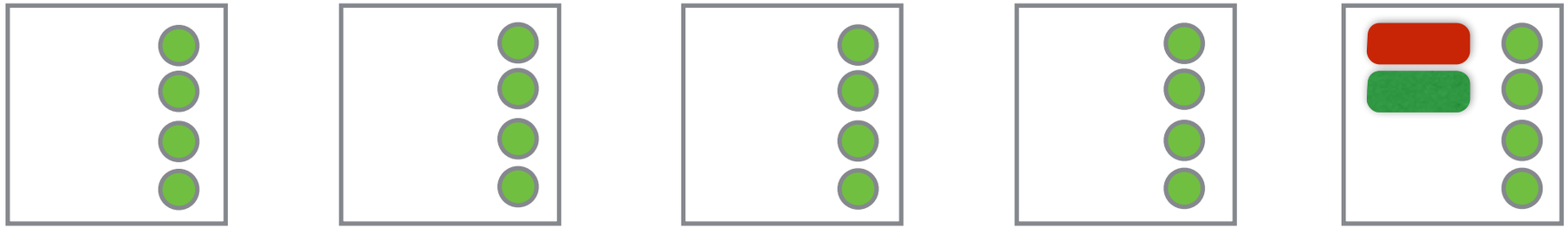


student 1 student 2 student 3 student 4 student 5

page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

T = 12



- *12 time units to complete (mins)*
- *Each student requires 4 time units to complete*
- *Time between one student and the next finishing is 2 time units (mins)*

page 4
page 3
page 2
page 1

Group Reading - Pipeline with Two Sections

in words…

- S1 reads p1&2.
- S1 passes p1&2 to S2.
- S1 reads p3&4, concurrently, S2 reads p1&2.
- S1 passes p3&4 to S2, concurrently, S2 passes p1&2 to S3.
- S2 reads p3&4, S3 reads p1&2.
- S2 passes p3&4 to S3, S3 passes p1&2 to S4.
- S3 reads p3&4, S4 reads p1&2.
- S3 passes p3&4 to S4, S4 passes p1&2 to S5.
- S4 reads p3&4, S5 reads p1&2.
- S4 passes p3&4 to S5.
- S5 reads p3&4.

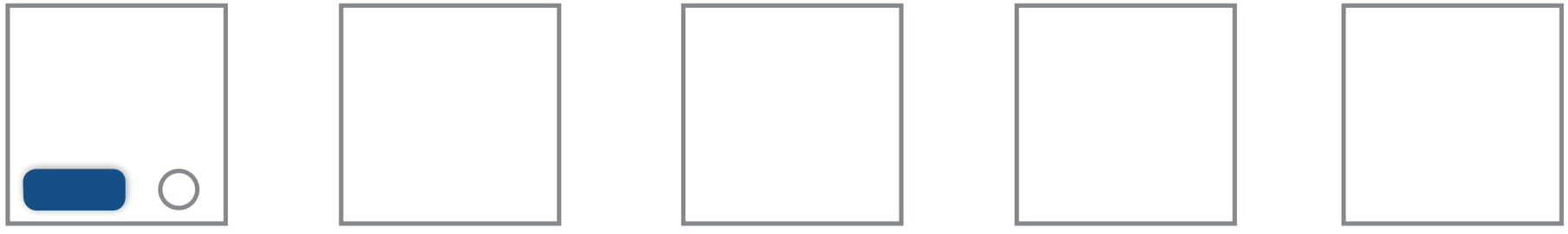
Figure 1: A step-by-step description of how five students (S1, S2, S3, S4, S5) read a four-page article partitioned into two 2-page parts (p1&2, p3&4) in a pipelined fashion.

How to speed this up?

- More of a good thing?

Group Reading - Pipeline with Four Sections

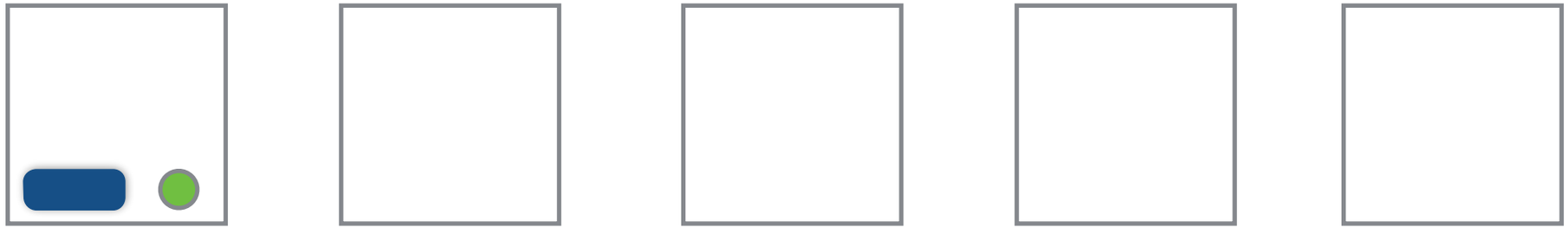
T = 0



page 4
page 3
page 2
page 1

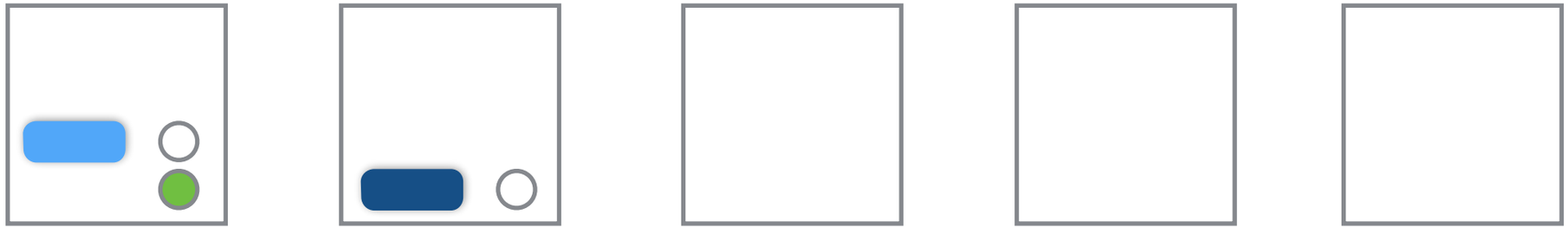
Group Reading - Pipeline with Four Sections

T = I



Group Reading - Pipeline with Four Sections

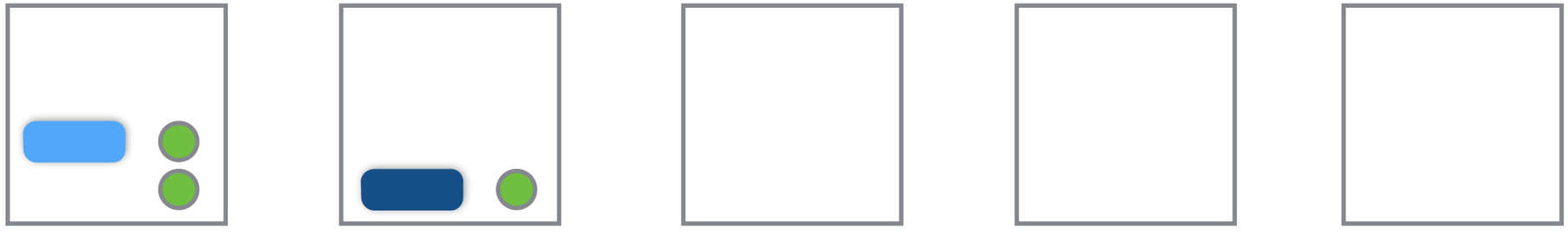
T = 2



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

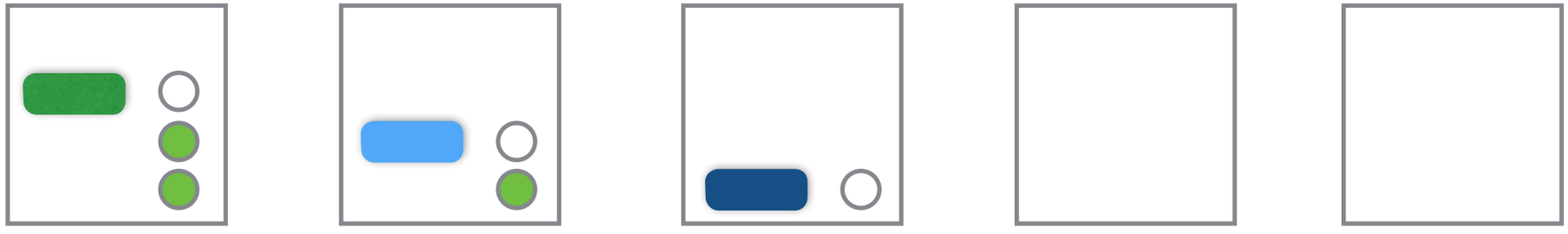
T = 2



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

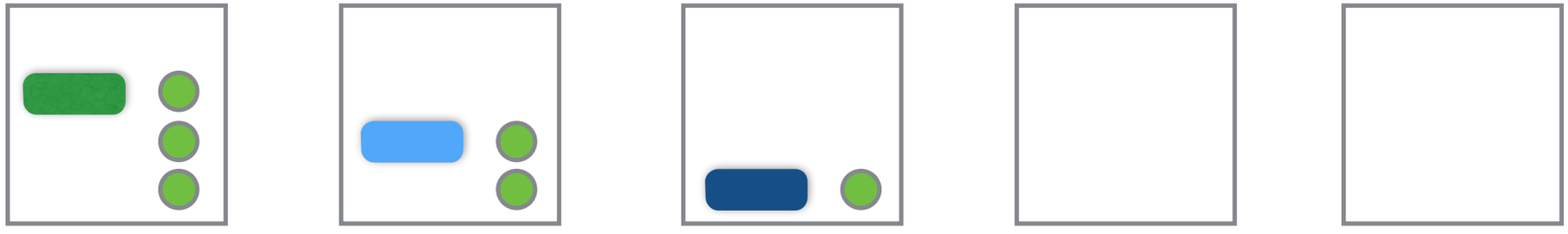
T = 3



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

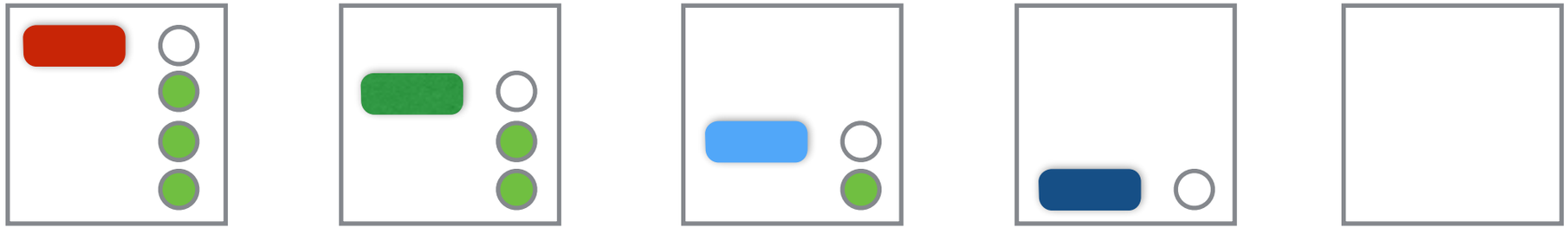
T = 3



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

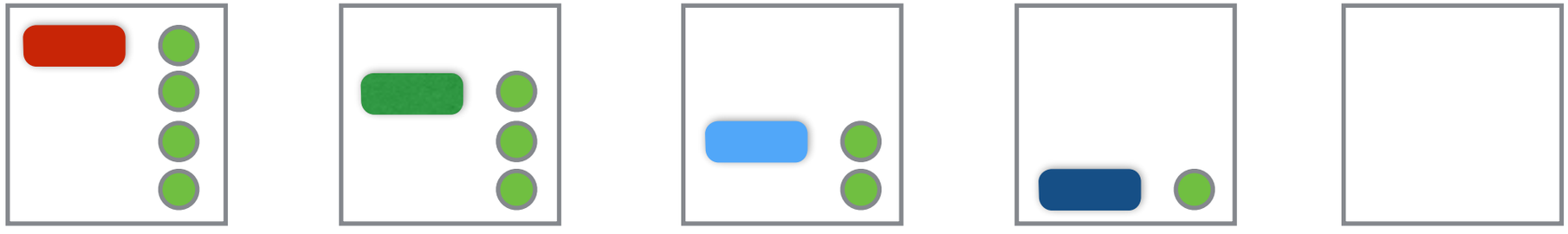
T = 4



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

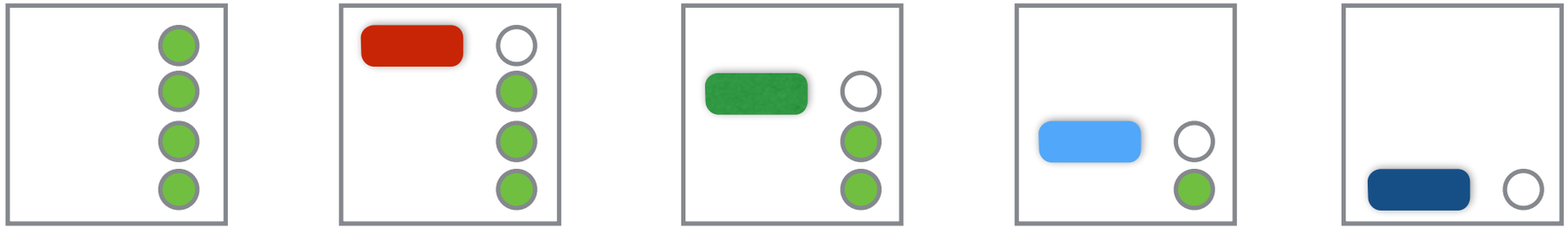
T = 4



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

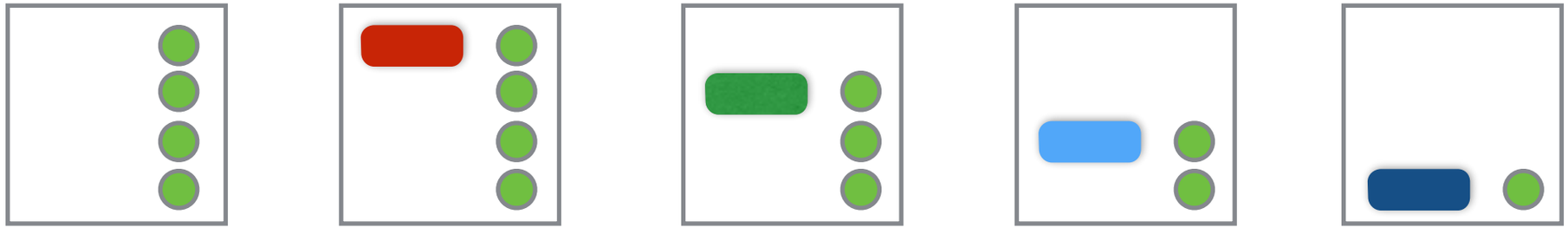
T = 5



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

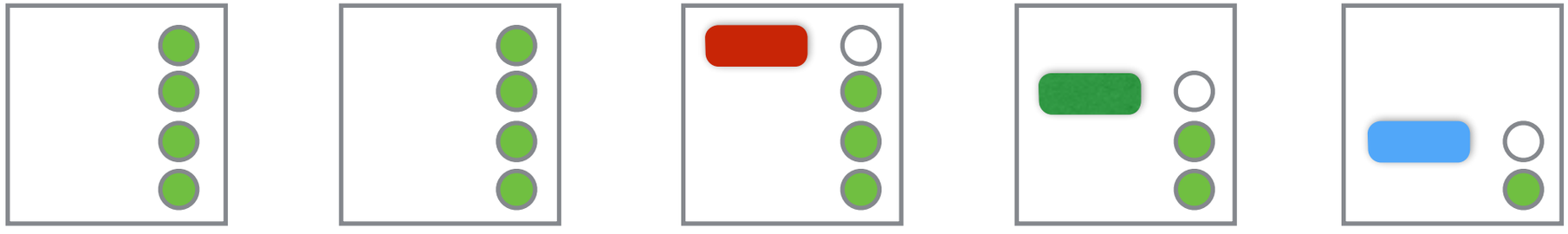
T = 5



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

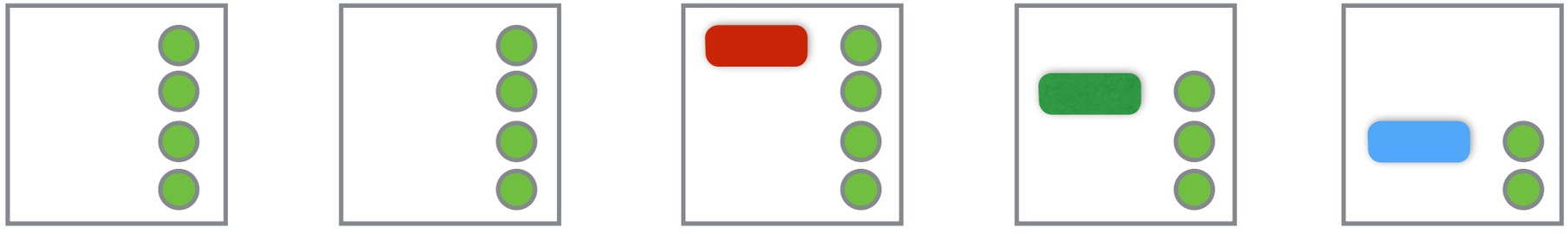
T = 6



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

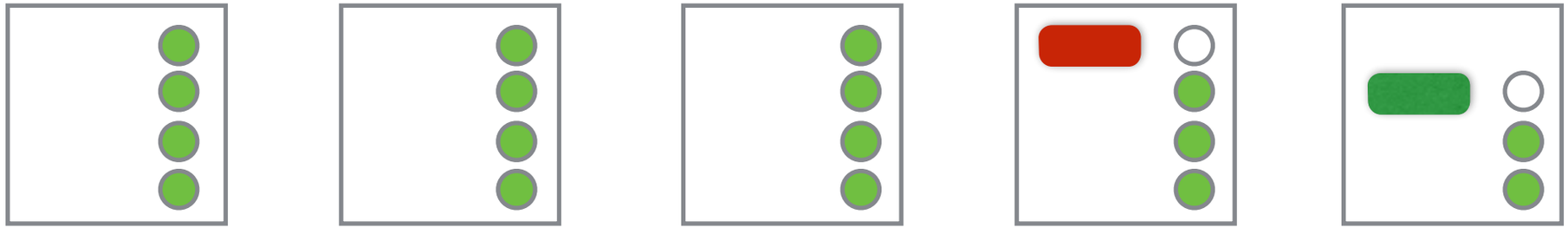
T = 6



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

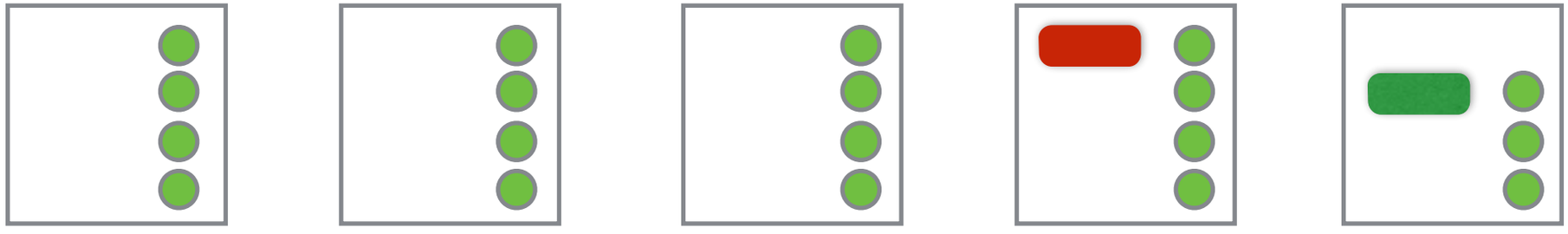
T = 7



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

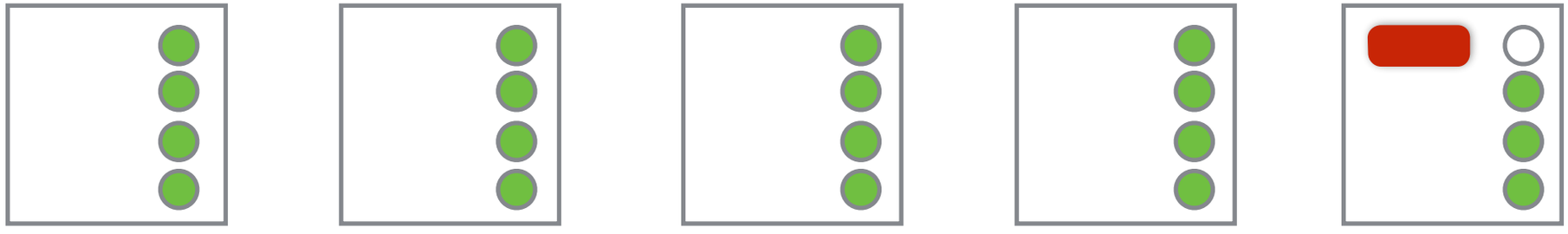
T = 7



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

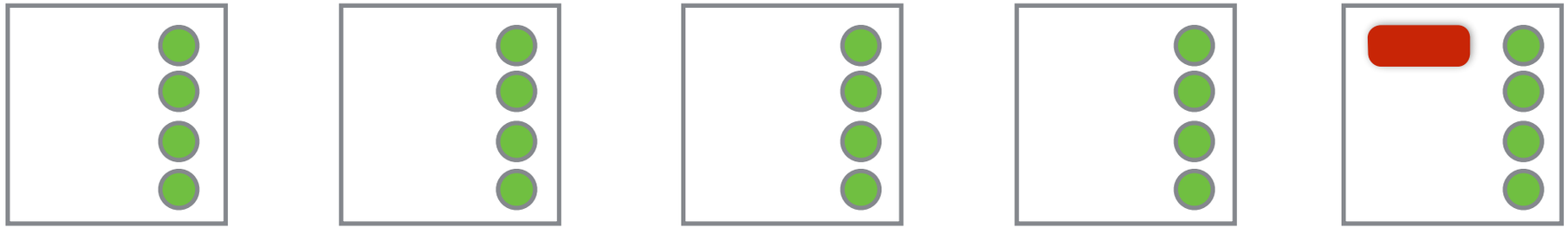
T = 8



page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

T = 8



- *8 time units to complete (mins)*
- *Each student requires 4 time units to complete*
- *Time between one student and the next finishing is 1 time units (mins)*

page 4
page 3
page 2
page 1

Group Reading - Pipeline with Four Sections

in words…

- S1 reads p1.
- S1 passes p1 to S2.
- S1 reads p2, concurrently, S2 reads p1.
- S1 passes p2 to S2, concurrently, S2 passes p1 to S3.
- S1 reads p3, S2 reads p2, S3 reads p1.
- S1 passes p3 to S2, S2 passes p2 to S3, S3 passes p1 to S4.
- S1 reads p4, S2 reads p3, S3 reads p2, S4 reads p1.
- S1 passes p4 to S2, S2 passes p3 to S3, S3 passes p2 to S4, S4 passes p1 to S5.
- S2 reads p4, S3 reads p3, S4 reads p2, S5 reads p1.
- S2 passes p4 to S3, S3 passes p3 to S4, S4 passes p2 to S5.
- S3 reads p4, S4 reads p3, S5 reads p2.
- S3 passes p4 to S4, S4 passes p3 to S5.
- S4 reads p4, S5 reads p3.
- S4 passes p4 to S5.
- S5 reads p4.

Figure 3: A step-by-step description of how five students (S1, S2, S3, S4, S5) read a four page article that is partitioned into four 1-page parts (p1, p2, p3, p4) in a pipelined fashion.

Comparing these 3 Approaches

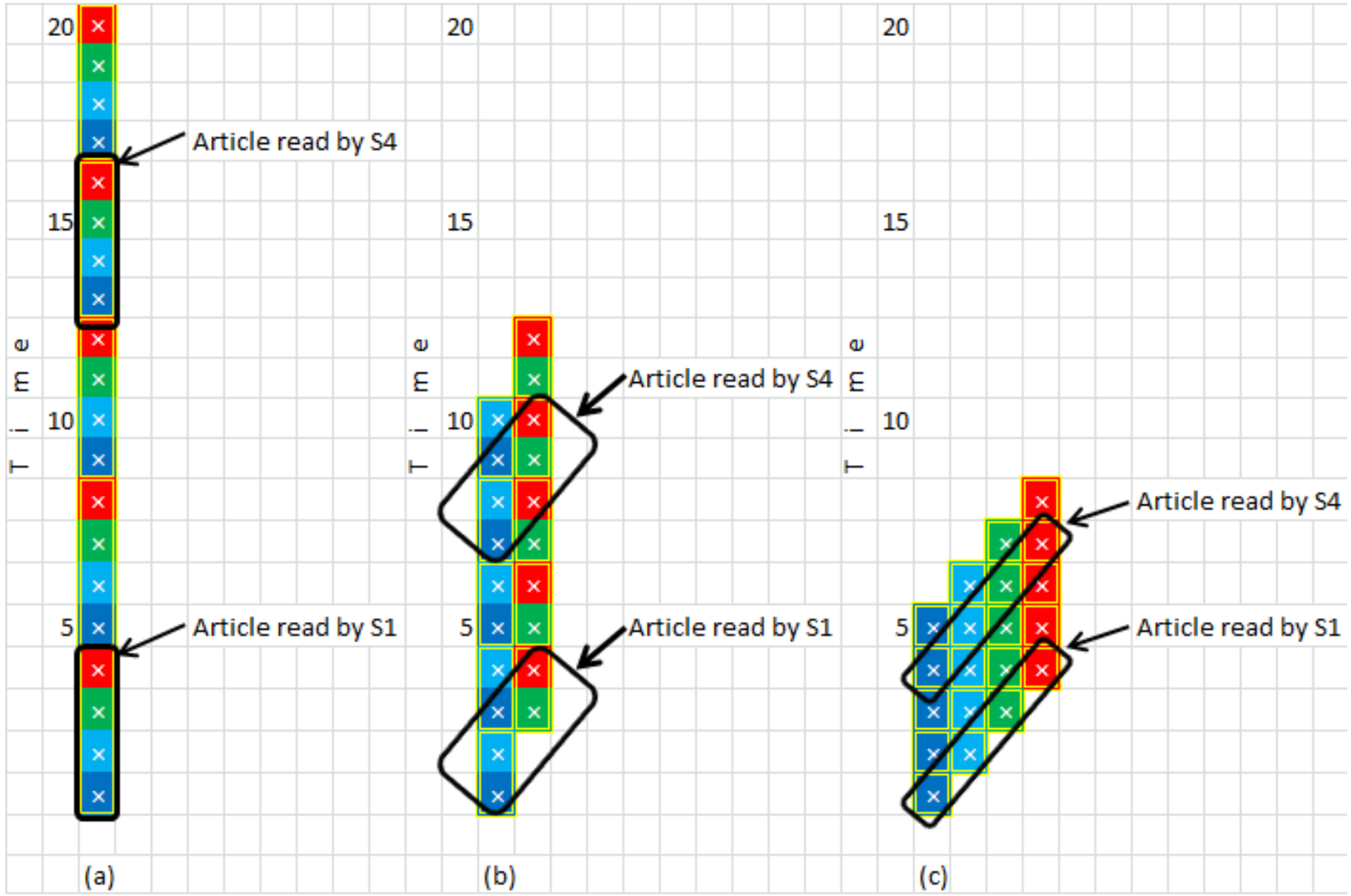


Figure 2: Five students reading one copy of a four-page article at the rate of 1 page per minute. Each box with four squares, two squares, or one square shows a part of the article, and each × indicates the page being read by a student at the particular time indicated by the y-axis. This figure assumes that each passing step takes zero time. (a) Full article read and then passed (non-pipelined, baseline for comparison); (b) article divided into two 2-page parts, each part read and then passed (pipelined); (c) article divided into four 1-page parts, each part read and then passed (pipelined, more extensively than (b)). (For scenarios in (b) and (c), Figure 1 and Figure 3 respectively describe which part is being read by which student at what time.)

How to speed this up?

- More of a good thing?

How to speed this up?

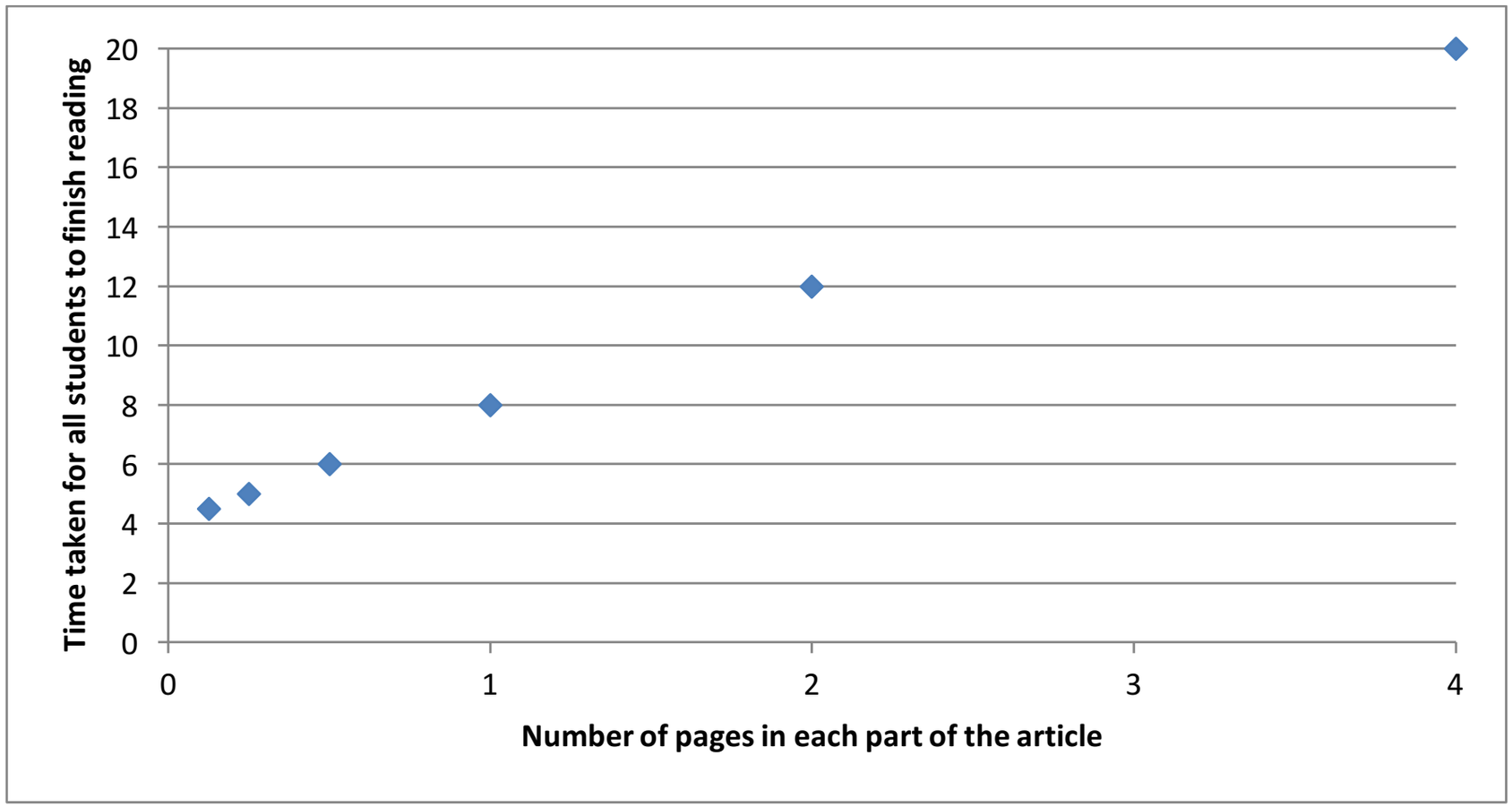


Figure 4: A comparison of total time for five students to finish reading a four page article, assuming each student reads one page per minute and that each passing step takes zero time, when the article is in one 4-

- Diminishing returns — less improvement as we go to smaller and smaller sections

Examples of Pipelining

Student inputs

- A class reading a book
- Dishwashing: wash, dry, put-away
- Assembly lines
- Microprocessor instruction processing (steps)
- Streaming over multiple network hops (Netflix)
- Graduation ceremony tasks
- Bucket brigade
- Blaze pizza — pizza ordering/production
- Optimized scheduling of workers/tasking a project or organization
    - supply chain management

Examples of Pipelining

Table 1: Correspondence between the above three example scenarios.

	Students reading an article	**Truck assembly line**	**Microprocessor processing instructions**
1	The article read by one student	One truck's assembly completed	One instruction processed
2	Partitioning an article into k parts	Partitioning the task of assembling a truck into k assembly steps	Partitioning the task of processing one instruction into k sub-tasks
3	A part of the article read by one student (P1, P2, …)	A stage of the assembly line works on one truck (A1, A2, …)	A hardware module performs a sub-task for one instruction (M1, M2, …)
4	A_i: i^{th} student completes reading the article	T_i: Assembly of i^{th} truck completed	I_i: Processing of i^{th} instruction completed
5	A_{ij}: i^{th} student completes reading the j^{th} part of the article	T_{ij}: j^{th} step in the assembly of i^{th} truck completed	I_{ij}: j^{th} sub-task in processing of i^{th} instruction completed

Overview

- Explore the concept of pipelining through an example
  - Group reading
  - Some reality checks
    - Refining our models
    - Limitations of pipelinings
- Characterizing general pipelining approaches

Discussion: Limitations of Pipelining/Model

- *What has our model neglected?*
- *What can go wrong or what are costs associated with this speed-up?*

Discussion: Limitations of Pipelining/Model

Student inputs

- *Assumed each student read at the speed*
  - *Limited by your slowest processing stage (e.g., slowest reader)*
  - *Single point of failure*
- *Need multiple processing resources*
- *Didn't count transition time (e.g., time to hand pages to neighbor)*
- *Cost of ripping the book into pieces*
  - *How to build a car in 20 stages? Break an instruction into many sub-steps?*
- *Coordination or time synchronization*
- *Cost to flush a long pipeline — switching costs.*

Finite Transition Delay

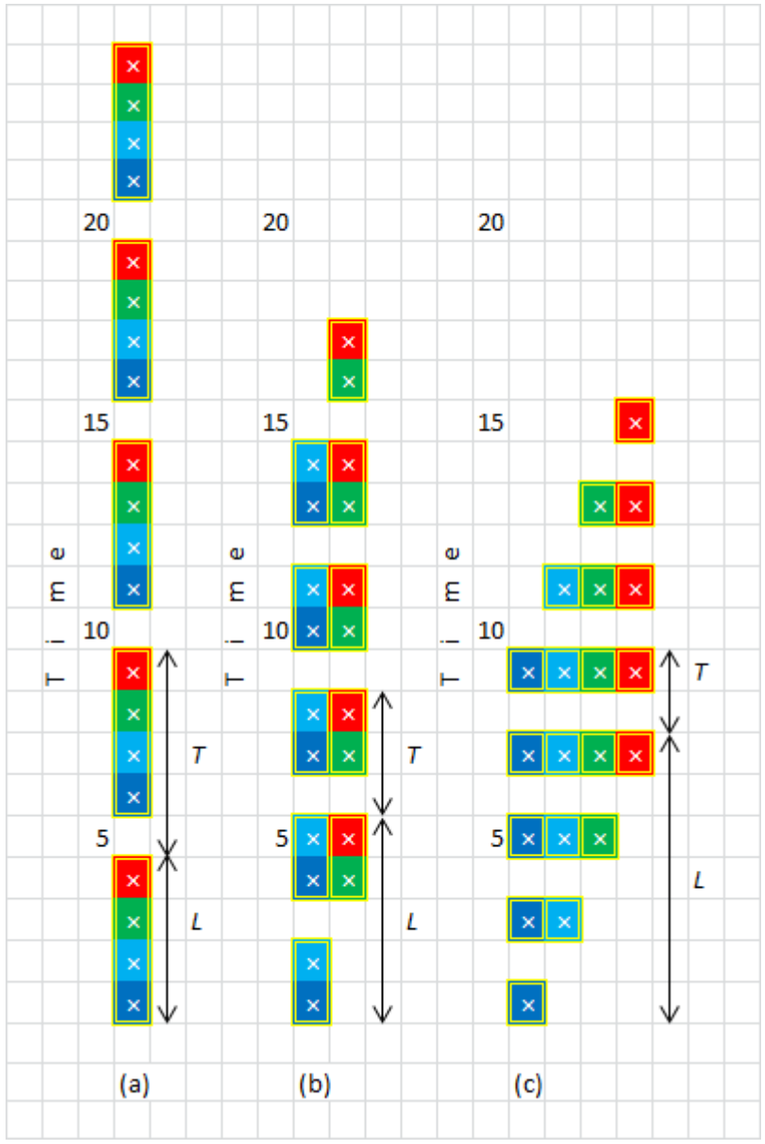


Figure 5: Five students reading one copy of a four-page article at the rate of 1 page per minute, where each passing step has overhead of $\delta = 1$ minute. Each box with four squares, two squares, or one square shows a part of the article, and each × indicates the page being read by a student at the particular time indicated by the y-axis. The empty boxes between reading steps are due to the 1 minute time overhead of each passing step. (a) Full article read and then passed (non-pipelined, baseline for comparison); (b) article divided into two two-page parts, each part read and then passed (pipelined); (c) article divided into four one-page parts, each part read and then passed (pipelined, more extensively than (b)). (For scenarios in (b) and (c), Figure 1 and Figure 3 respectively describe which part is being read by which student at what time.)

Throughput vs Latency

- *X time units to complete (mins)*
- *Each student requires Y time units to complete*
- *Time between one student and the next finishing is Z time units (mins)*

→ Latency

→ Throughput