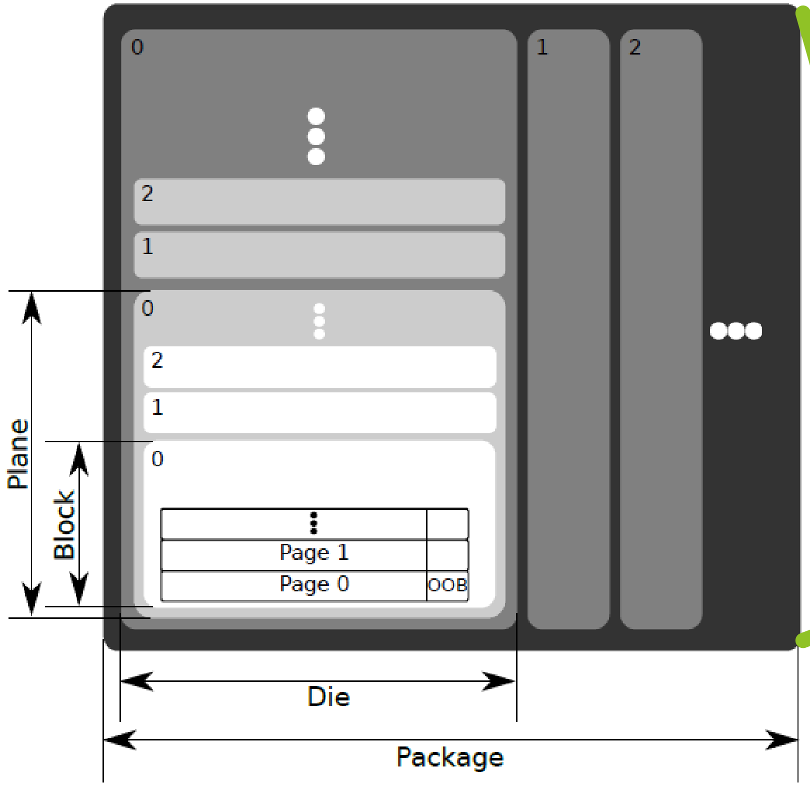
Need for a Deeper Cross-Layer Optimization for Dense NAND SSD to Improve Read Performance of Big Data Applications: A Case for Melded Pages

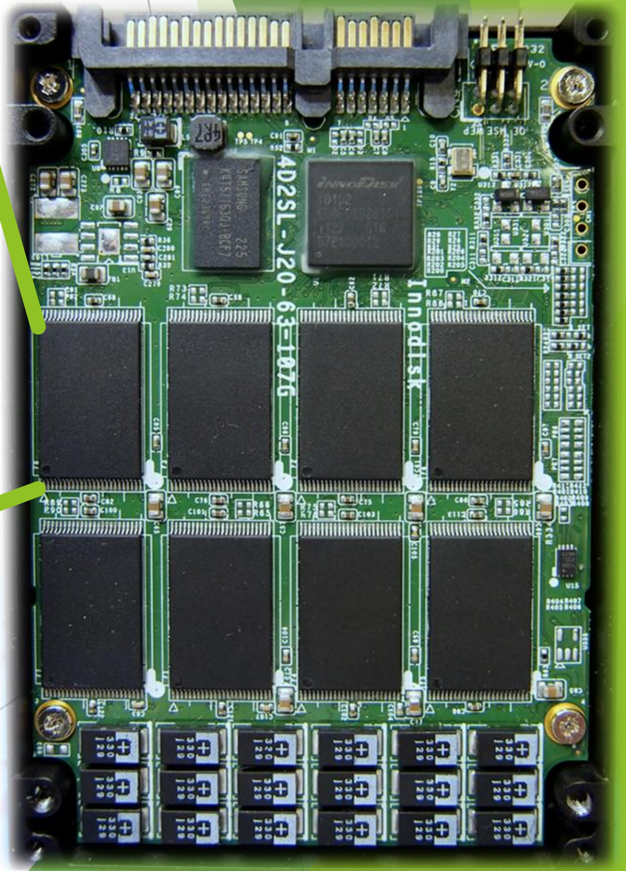
Arpith K, Indian Institute of Science, Bangalore

K. Gopinath, Indian Institute of Science, Bangalore

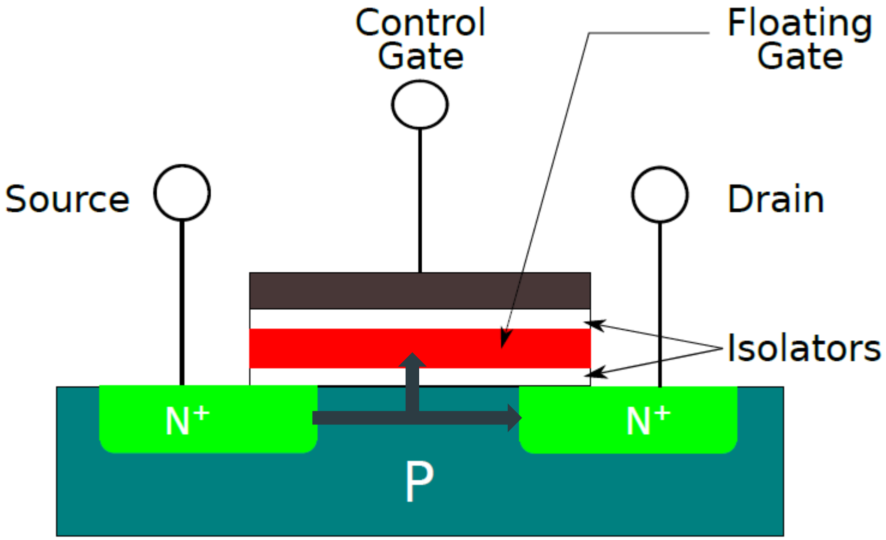
Organization of a Flash Packages

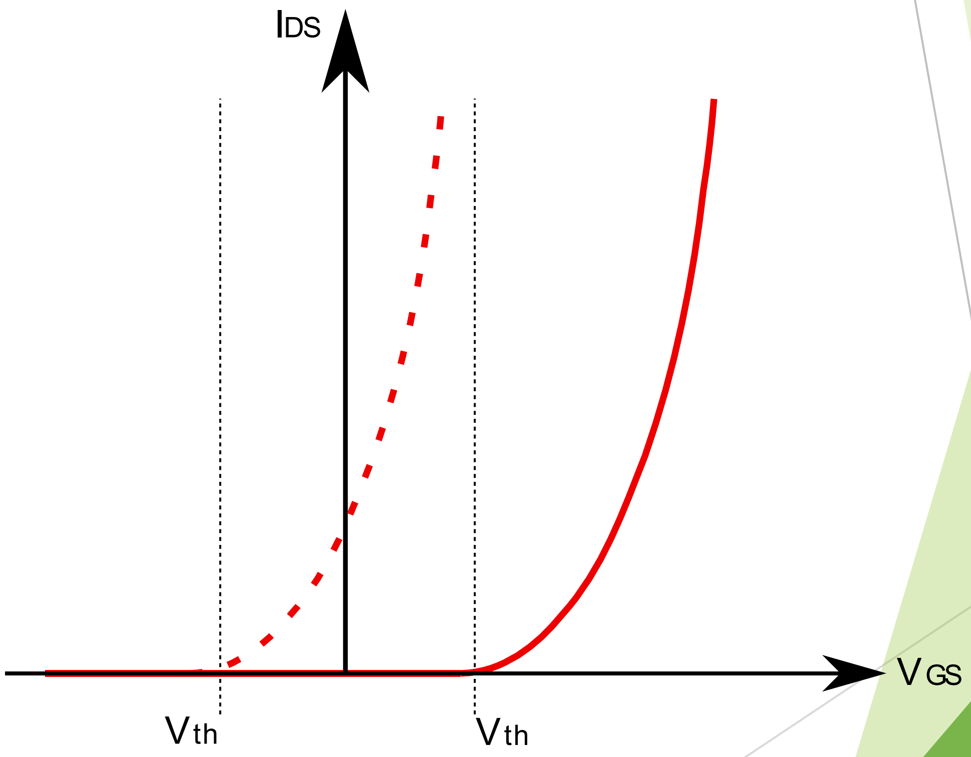
- **Die**
  - Smallest unit that can independently execute commands.
- **Plane**
  - Smallest unit to serve an I/O request in a parallel fashion.
- **Block**
  - Smallest unit that can be erased
- **Page**
  - Smallest unit that can be read or programed
- **Cell**



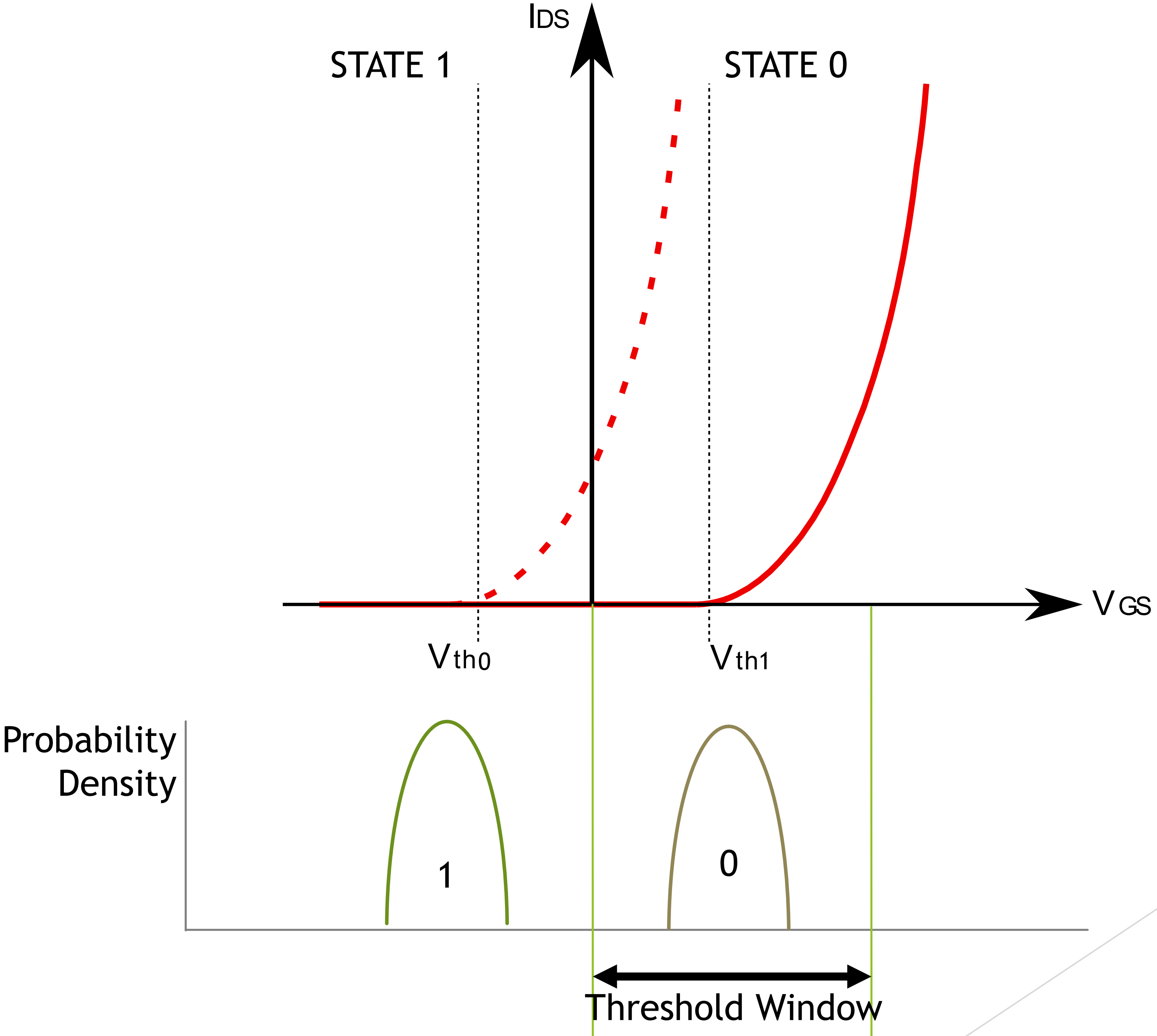


Floating Gate Transistors





- The presence of electrons in the floating gate increases the threshold voltage of the cell

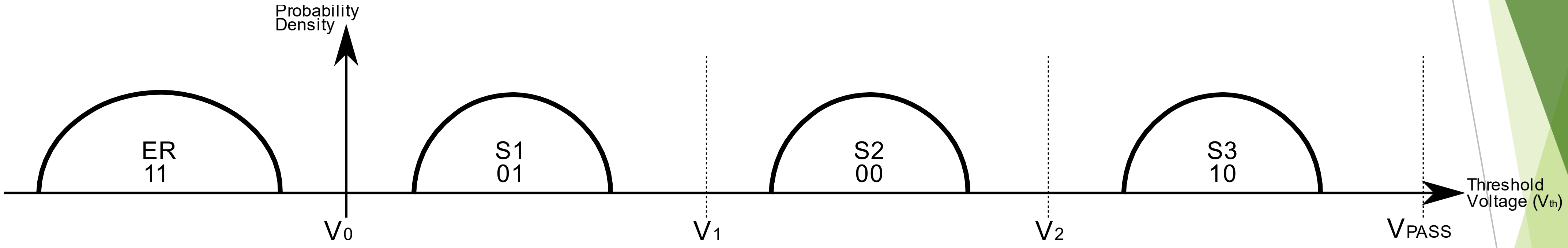


Threshold Voltage

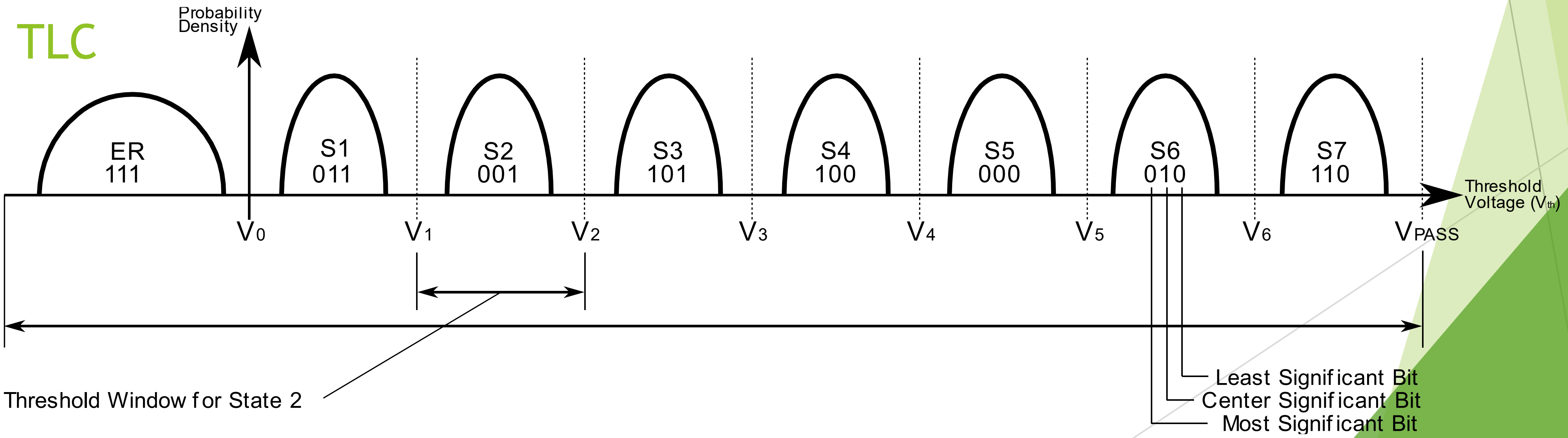
Reads

- Number of threshold voltage states determines how many bits a transistor can store.

MLC



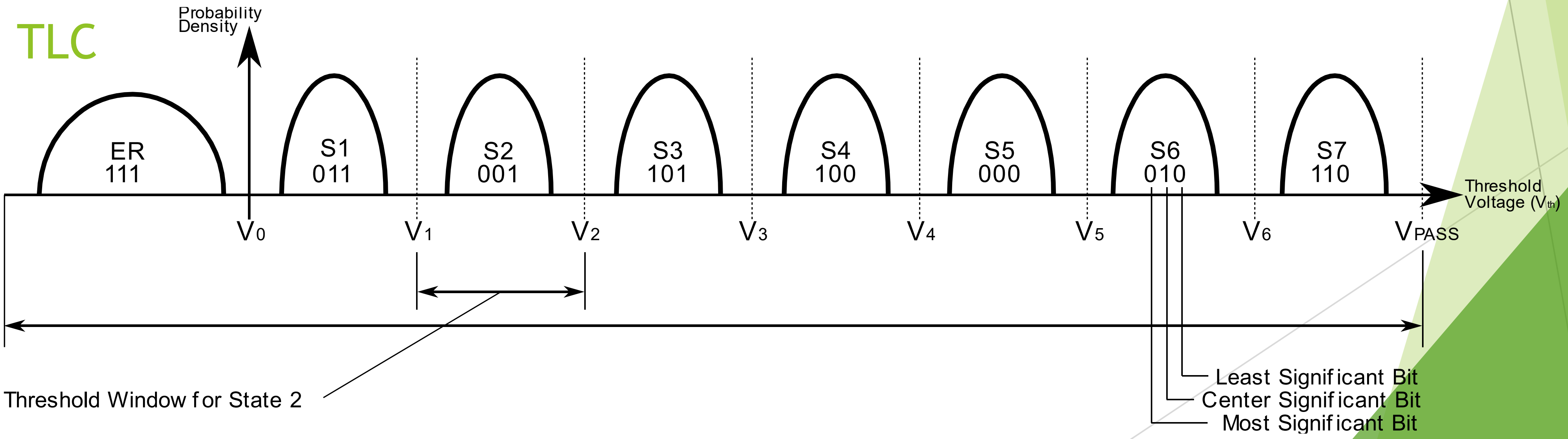
TLC



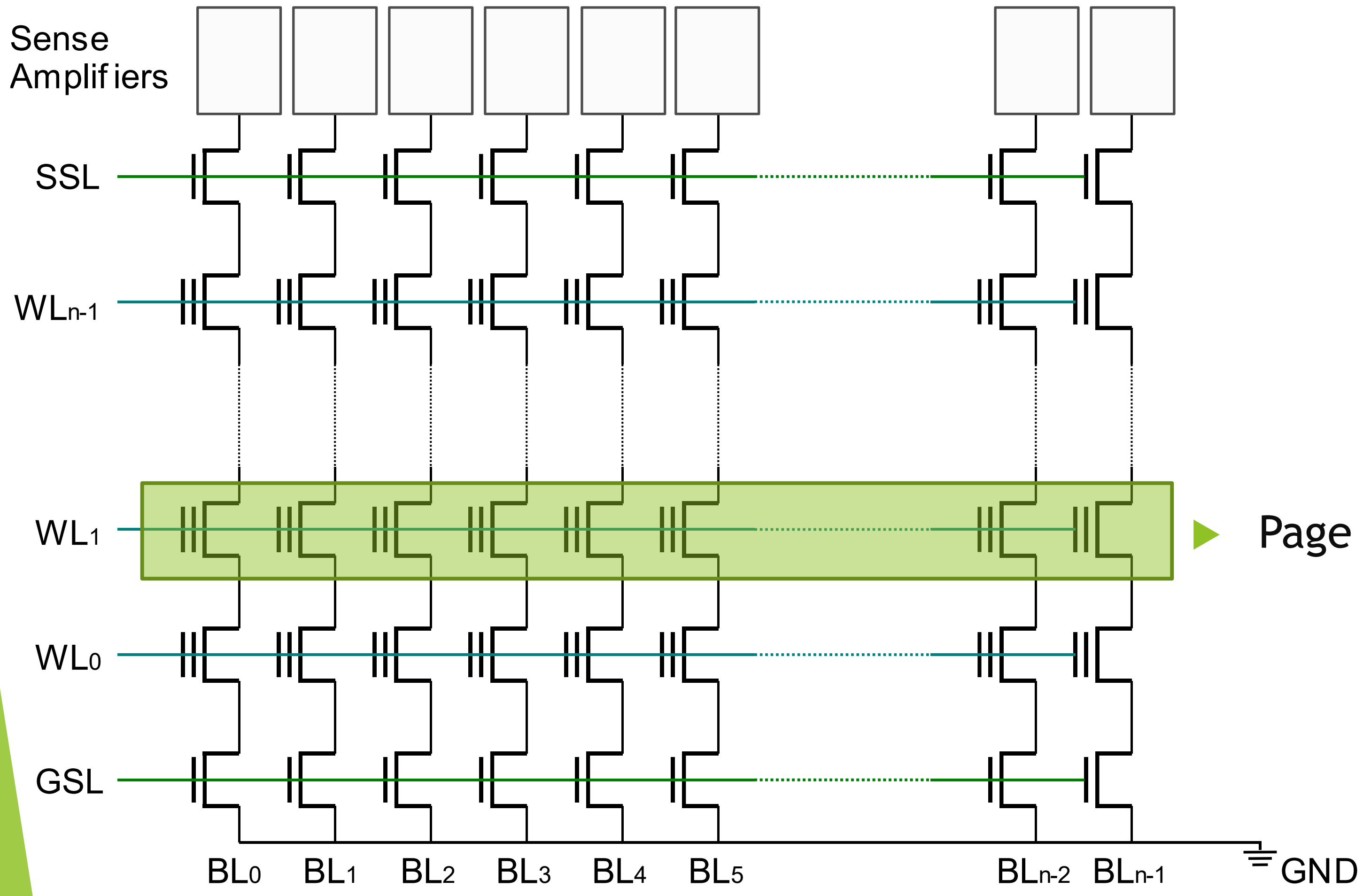
Reads

- LSB
  - V_3
- CSB
  - $V_{1,}$ V_5
- MSB
  - V_0, V_2, V_4, V_6

TLC

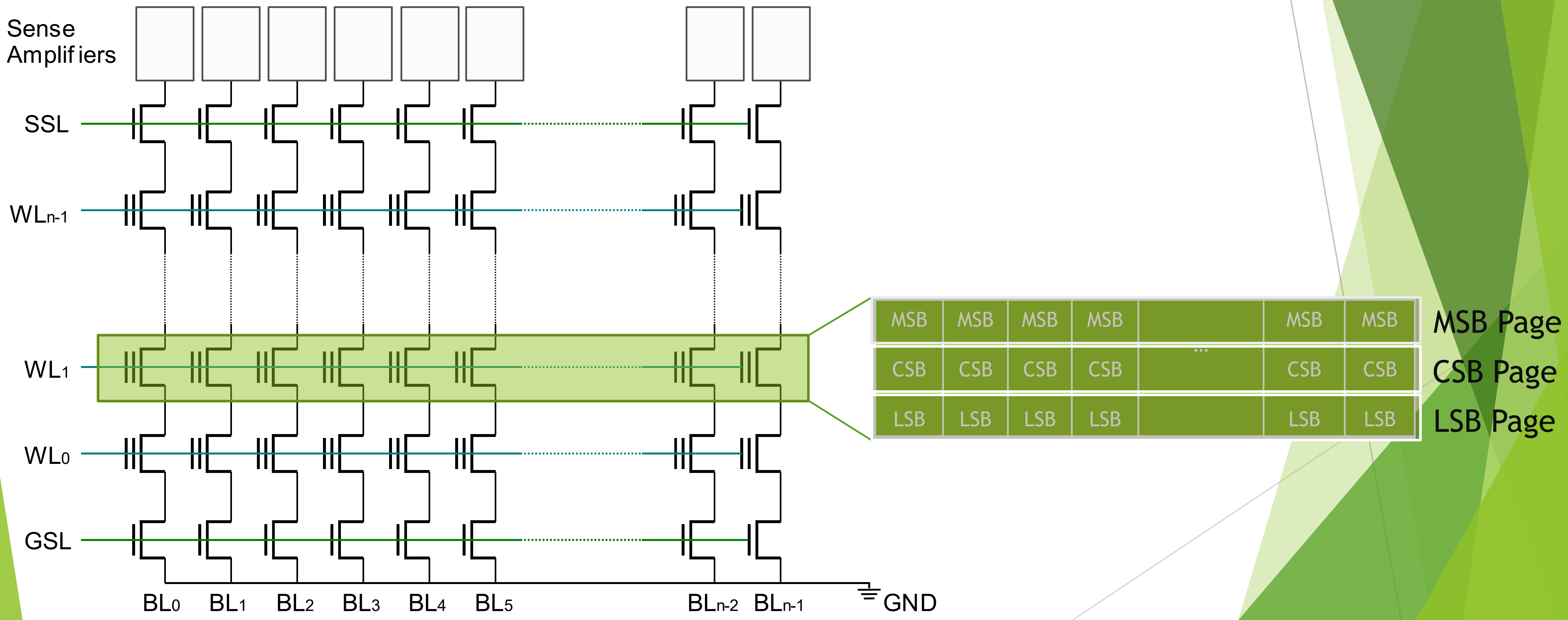


Organization of Transistors in a Block



- **Page** (Smallest unit that can be read or programed)

Organization of Transistors in a Block

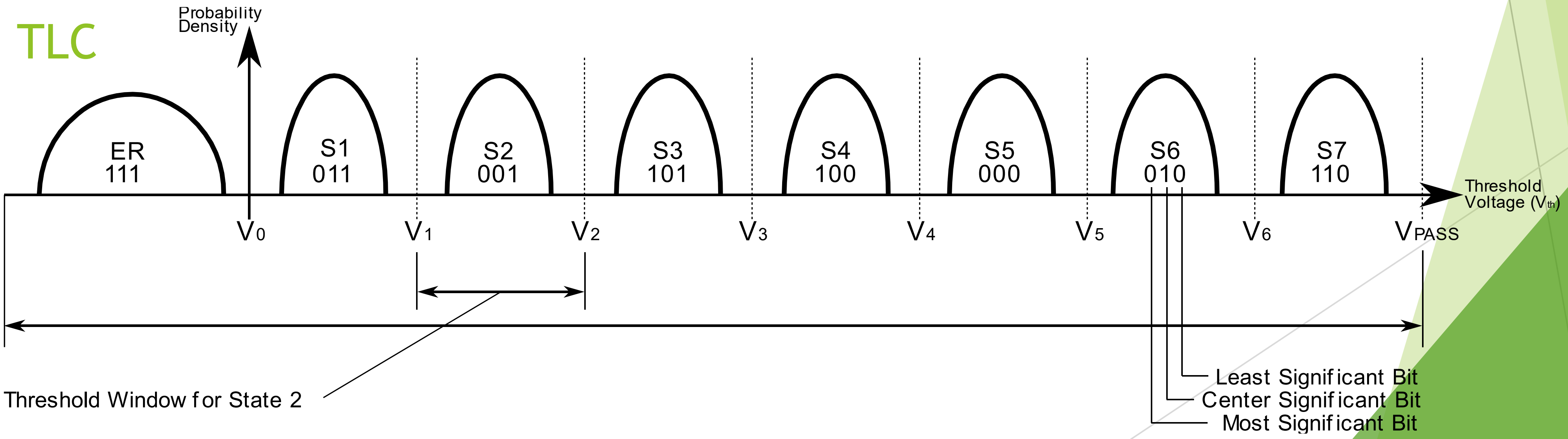


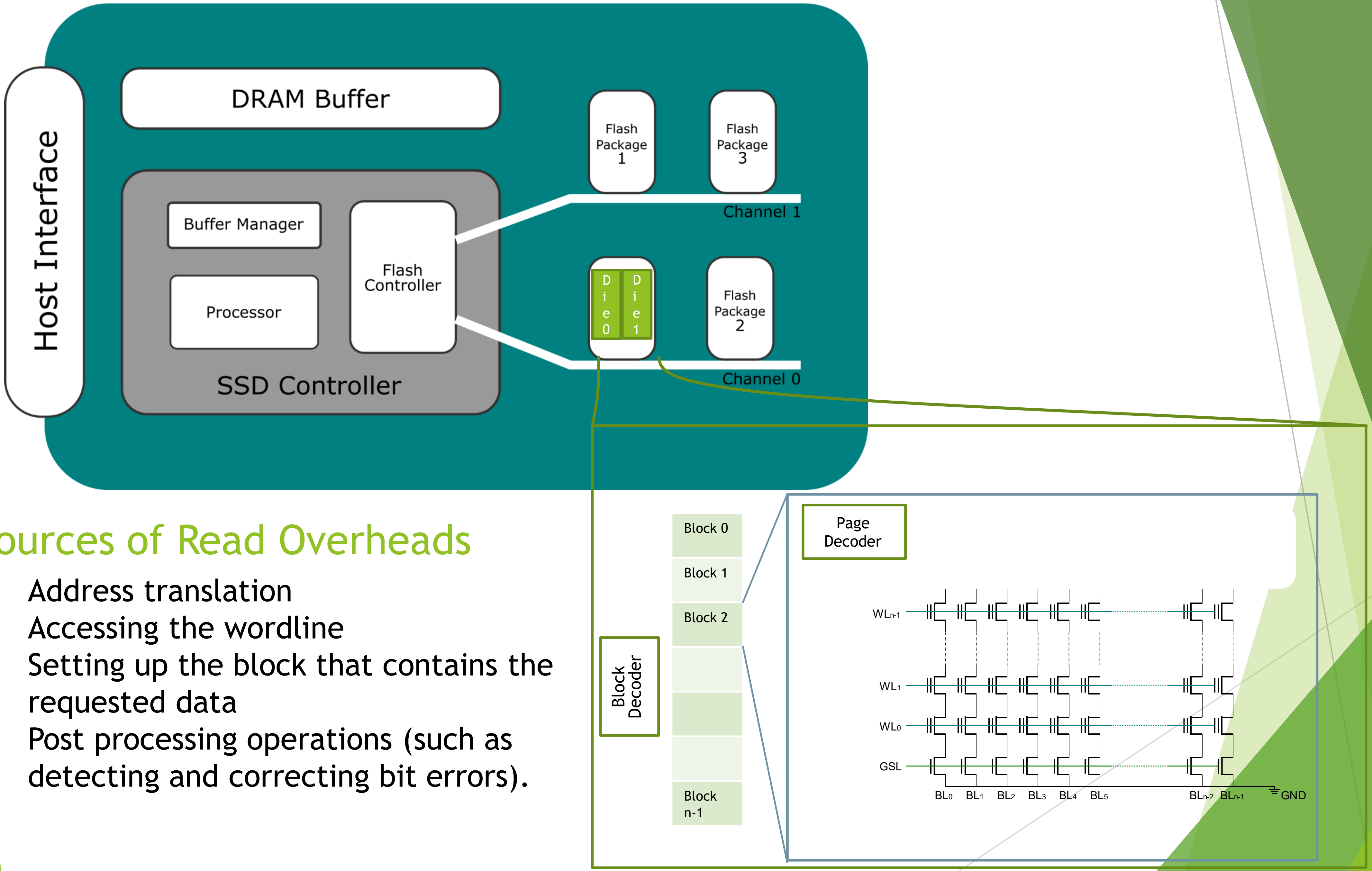
MSB	MSB	MSB	MSB		MSB	MSB	MSB Page
CSB	CSB	CSB	CSB	...	CSB	CSB	CSB Page
LSB	LSB	LSB	LSB		LSB	LSB	LSB Page

Reads Latency for TLC

Page	Latency (µs)
LSB Page	58
CSB Page	78
MSB Page	107

TLC

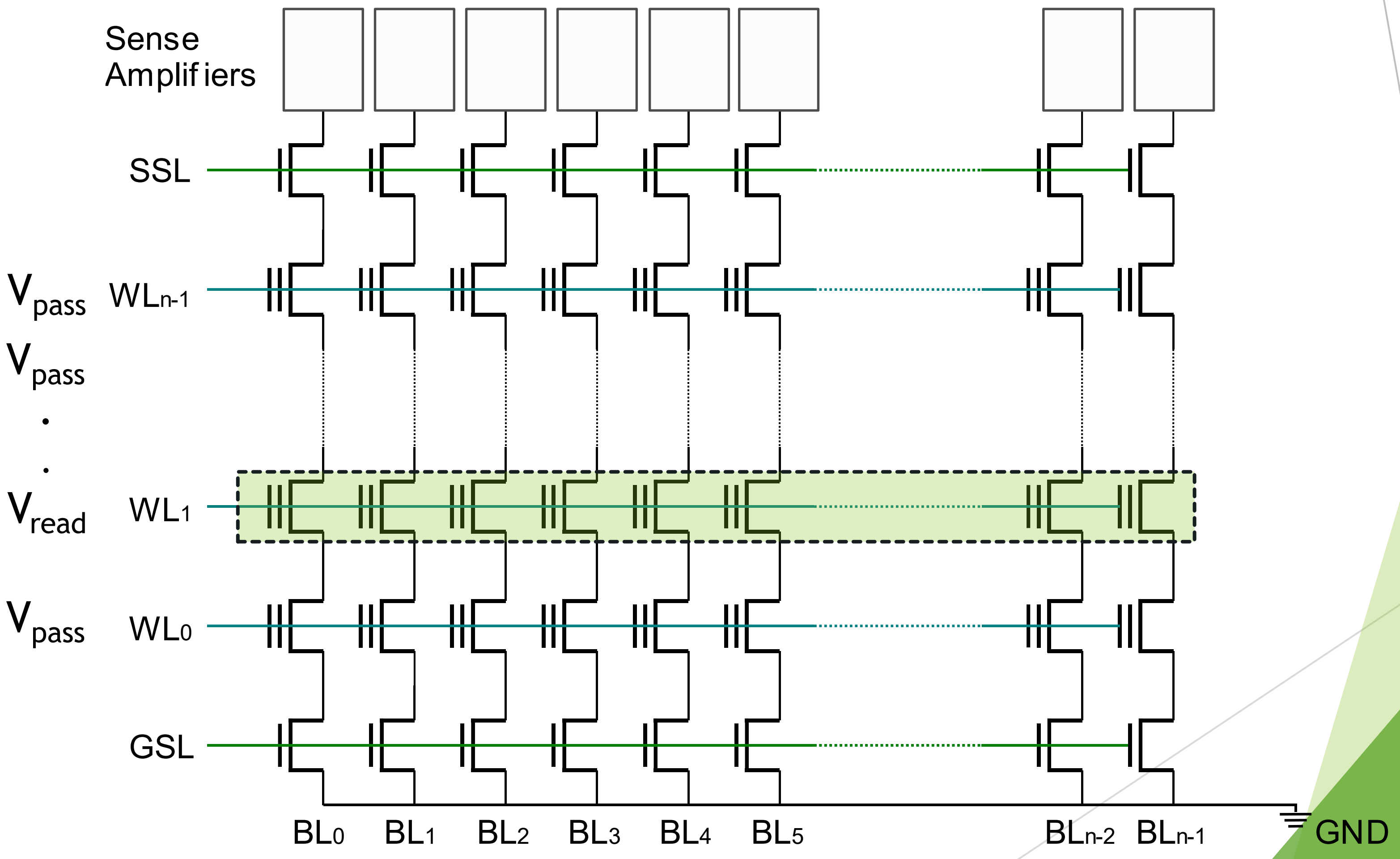


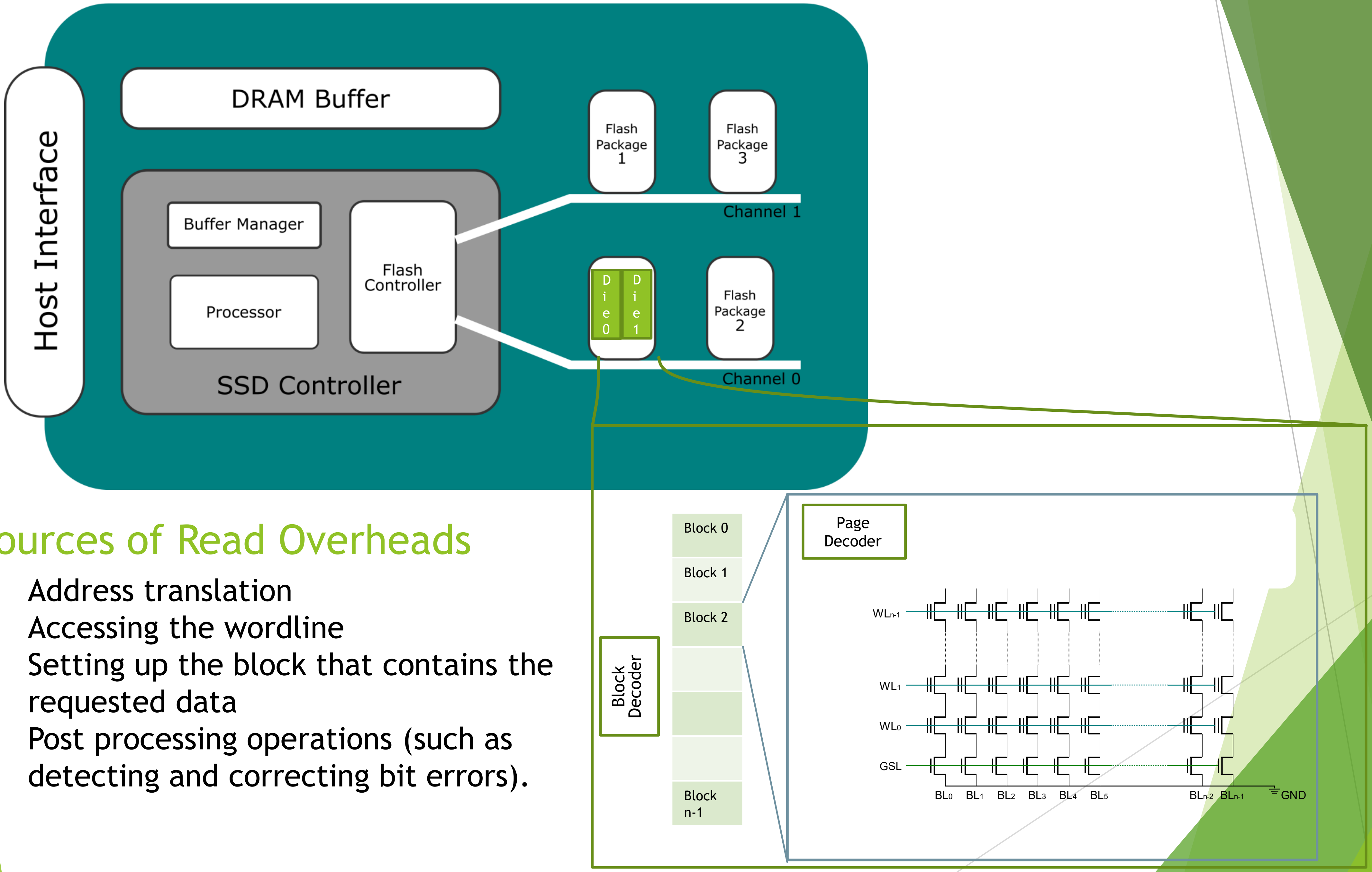


Sources of Read Overheads

- Address translation
- Accessing the wordline
- Setting up the block that contains the requested data
- Post processing operations (such as detecting and correcting bit errors).

Block Setup


Sense Amplifiers
SSL
V_pass WL_n-1
V_pass
V_read WL_1
V_pass WL_0
GSL
BL0
BL1
BL2
BL3
BL4
BL5
BLn-2
BLn-1
GND



Sources of Read Overheads

- Address translation
- Accessing the wordline
- Setting up the block that contains the requested data
- Post processing operations (such as detecting and correcting bit errors).

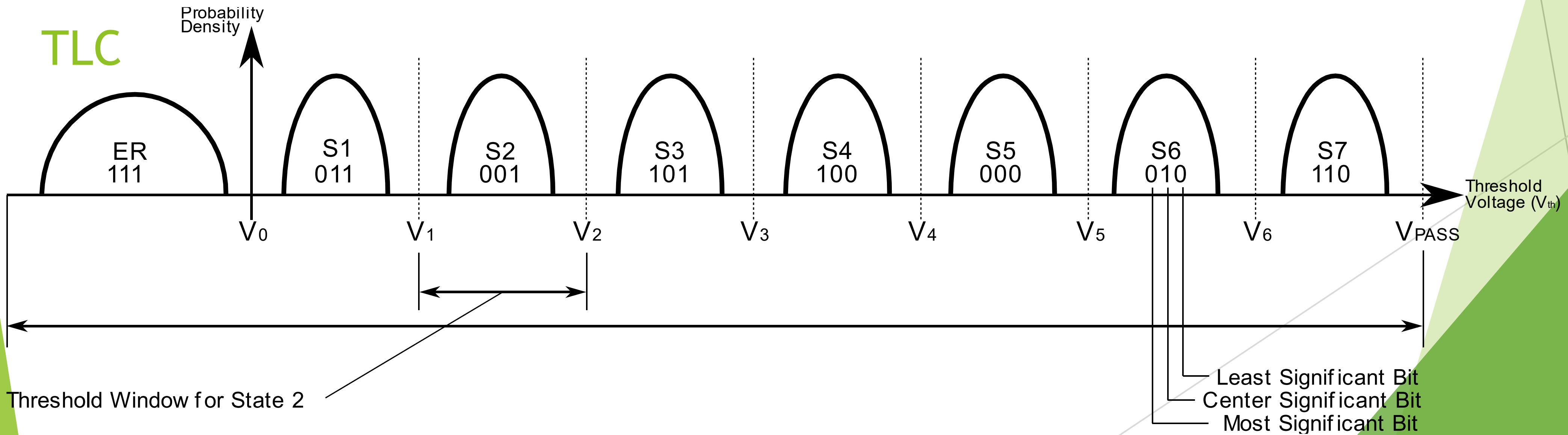
Reads

- X → Overhead. Includes time to address a wordline, apply pass through voltage (to other wordlines in that block) and post process data.
- Y → Time required to apply one read reference voltage and sense the cell’s conductivity.

Page	Latency (us)
LSB Page	58
CSB Page	78
MSB Page	107

- X + Y
- X + 2Y
- X + 4Y

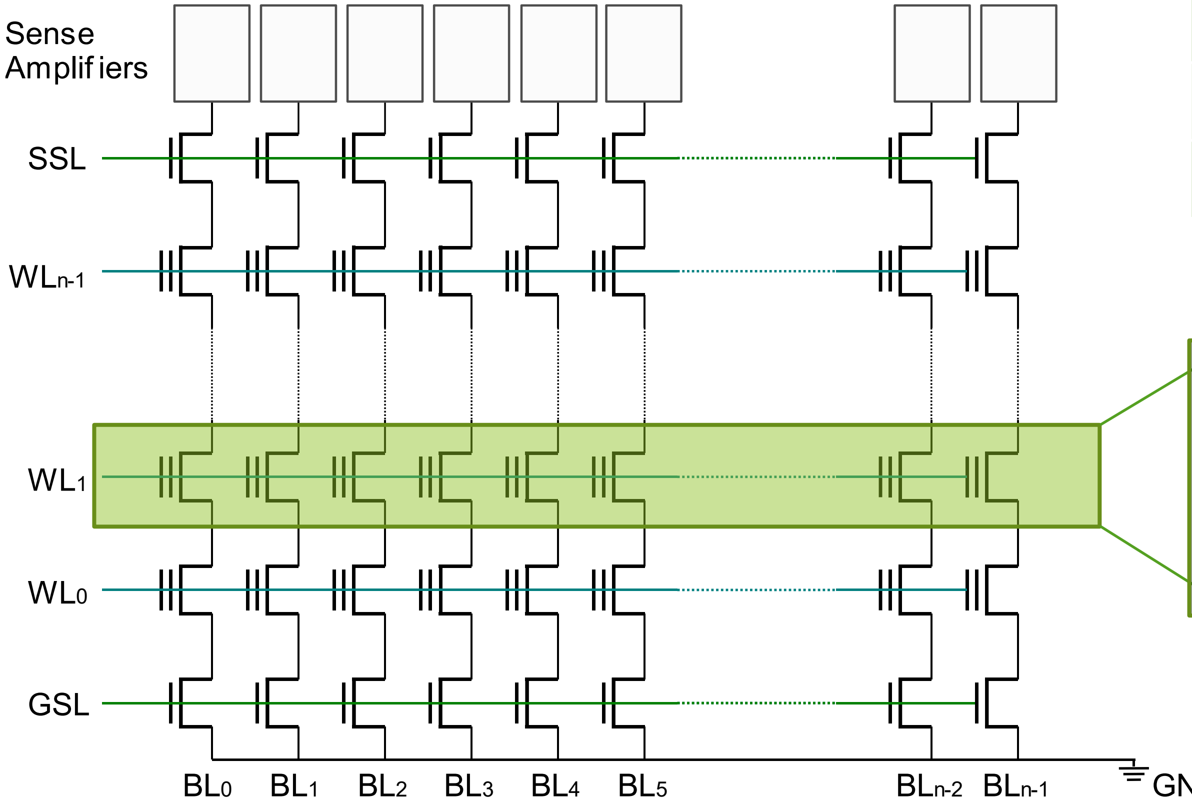
TLC

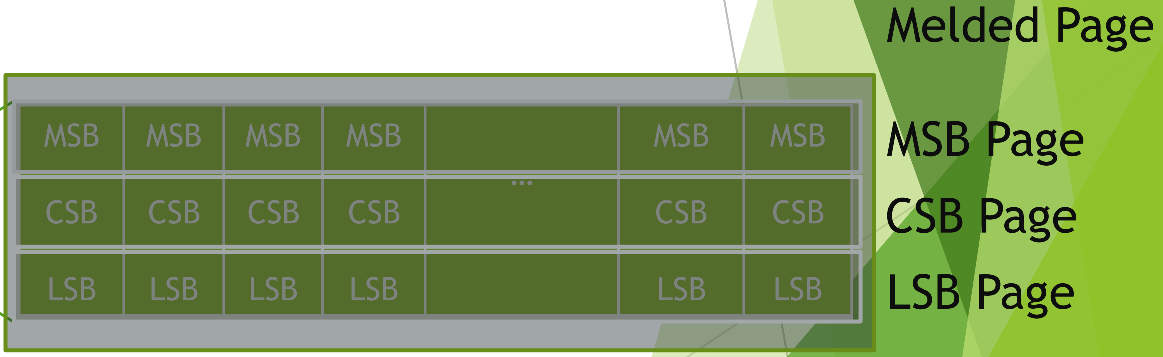


Meded-Pages

- Total time to read all three pages reduces from (3X + 7Y) to (X + 7Y)

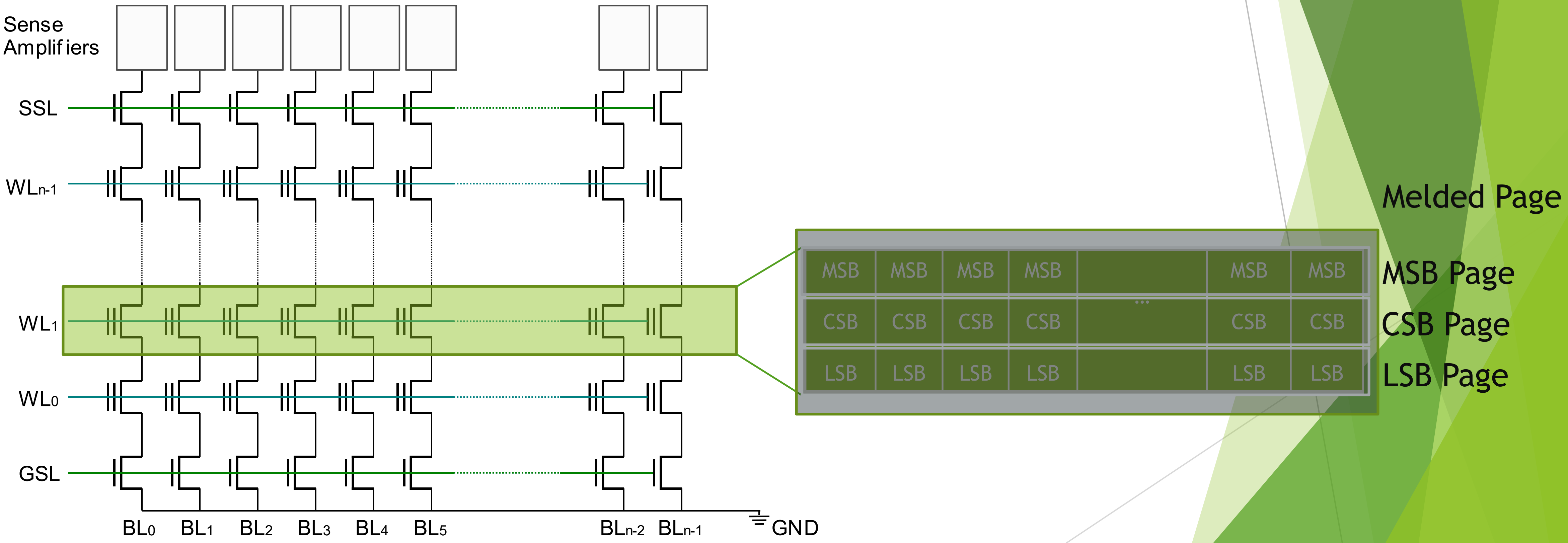
Page	Latency (us)	Latency MP (us)
LSB Page	58	166
CSB Page	78	
MSB Page	107	



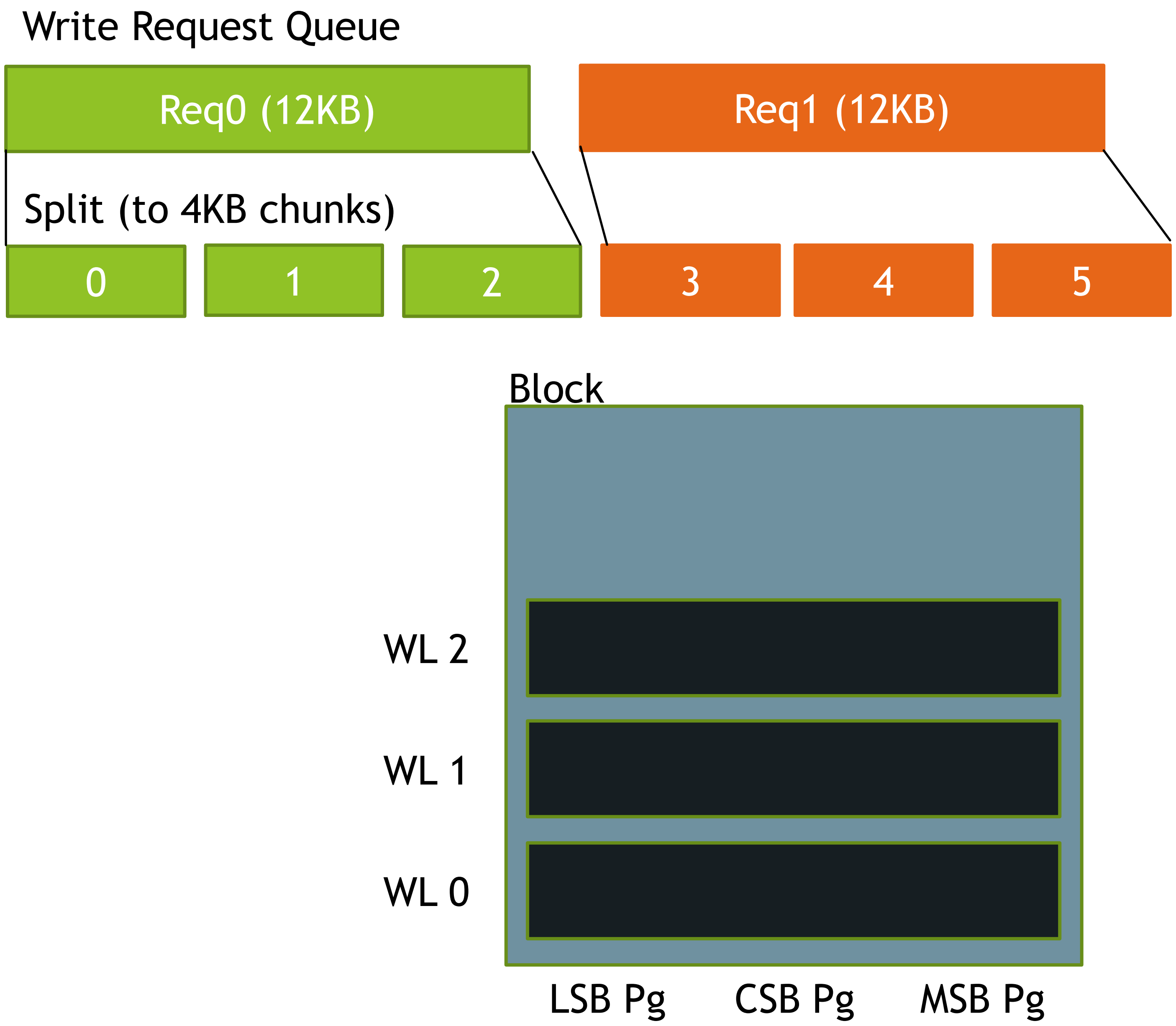


Meded-Pages

- Schedule the writes in such a way that, later, while reading, requests for data in LSB, CSB and MSB pages are all present in the read request queue.



Scheduling of Writes



Write Request Queue

Req0 (12KB)

Req1 (12KB)

Split (to 4KB chunks)

0

1

2

3

4

5

Block

WL 2

WL 1

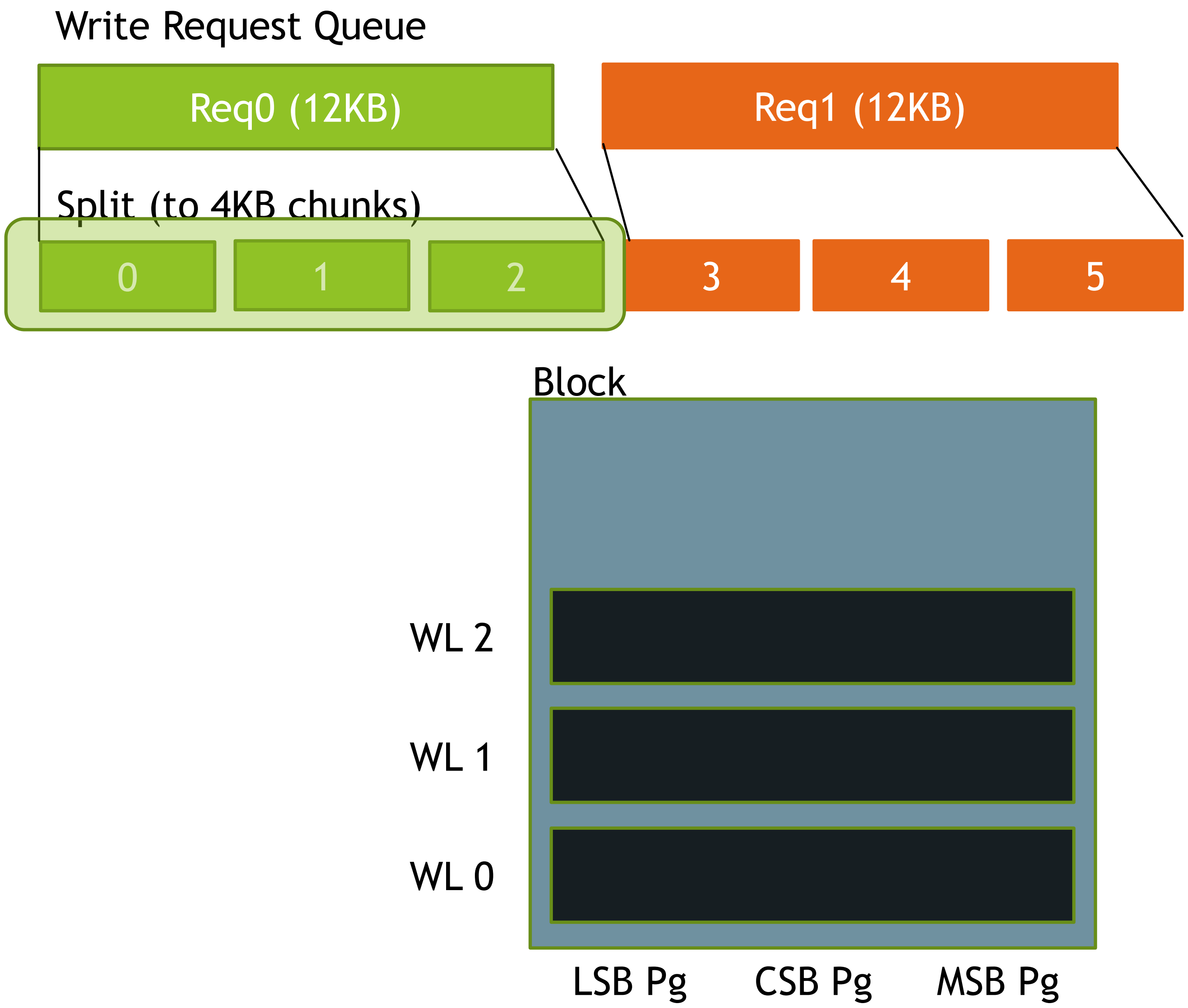
WL 0

LSB Pg

CSB Pg

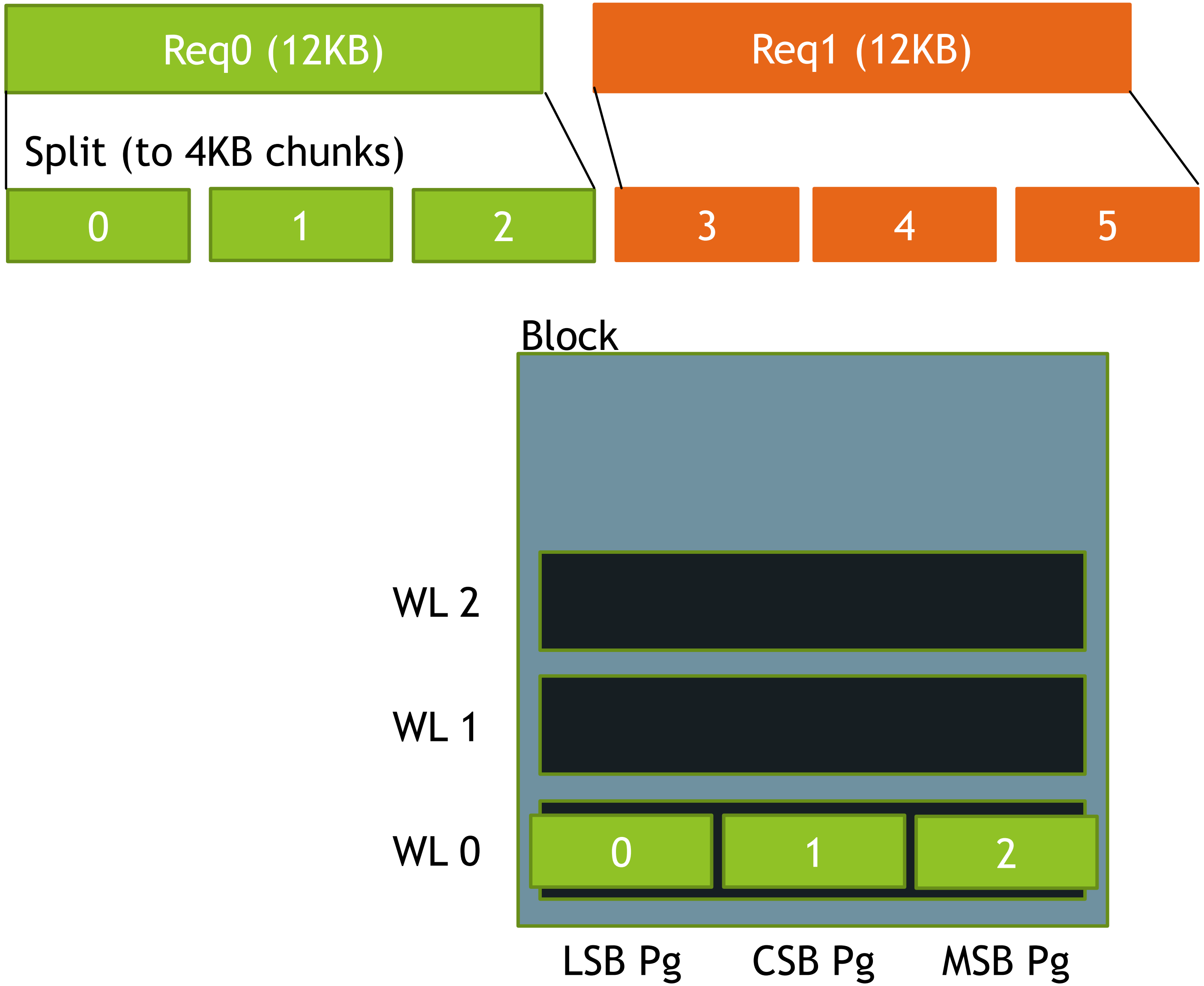
MSB Pg

Scheduling of Writes

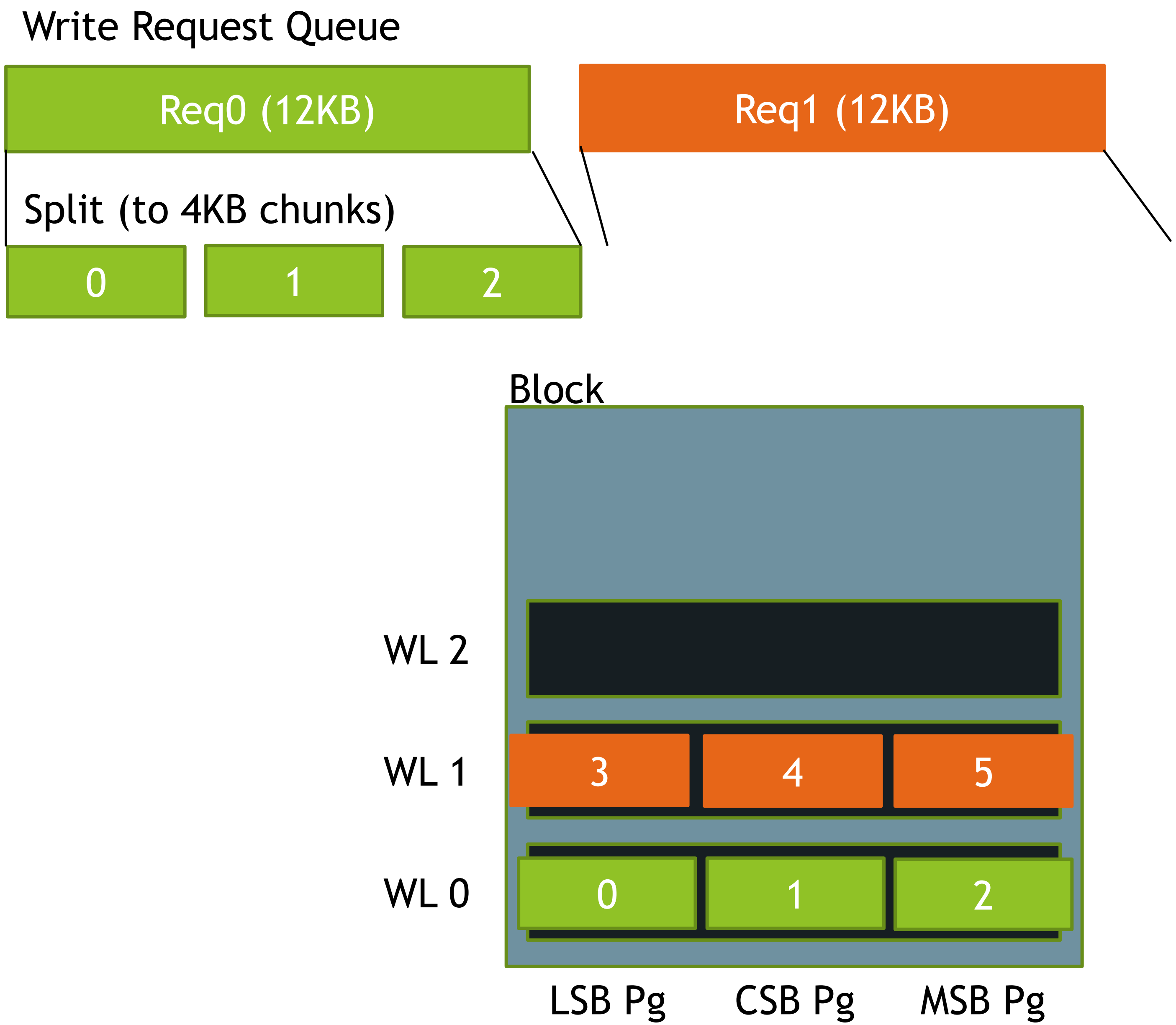


Scheduling of Writes

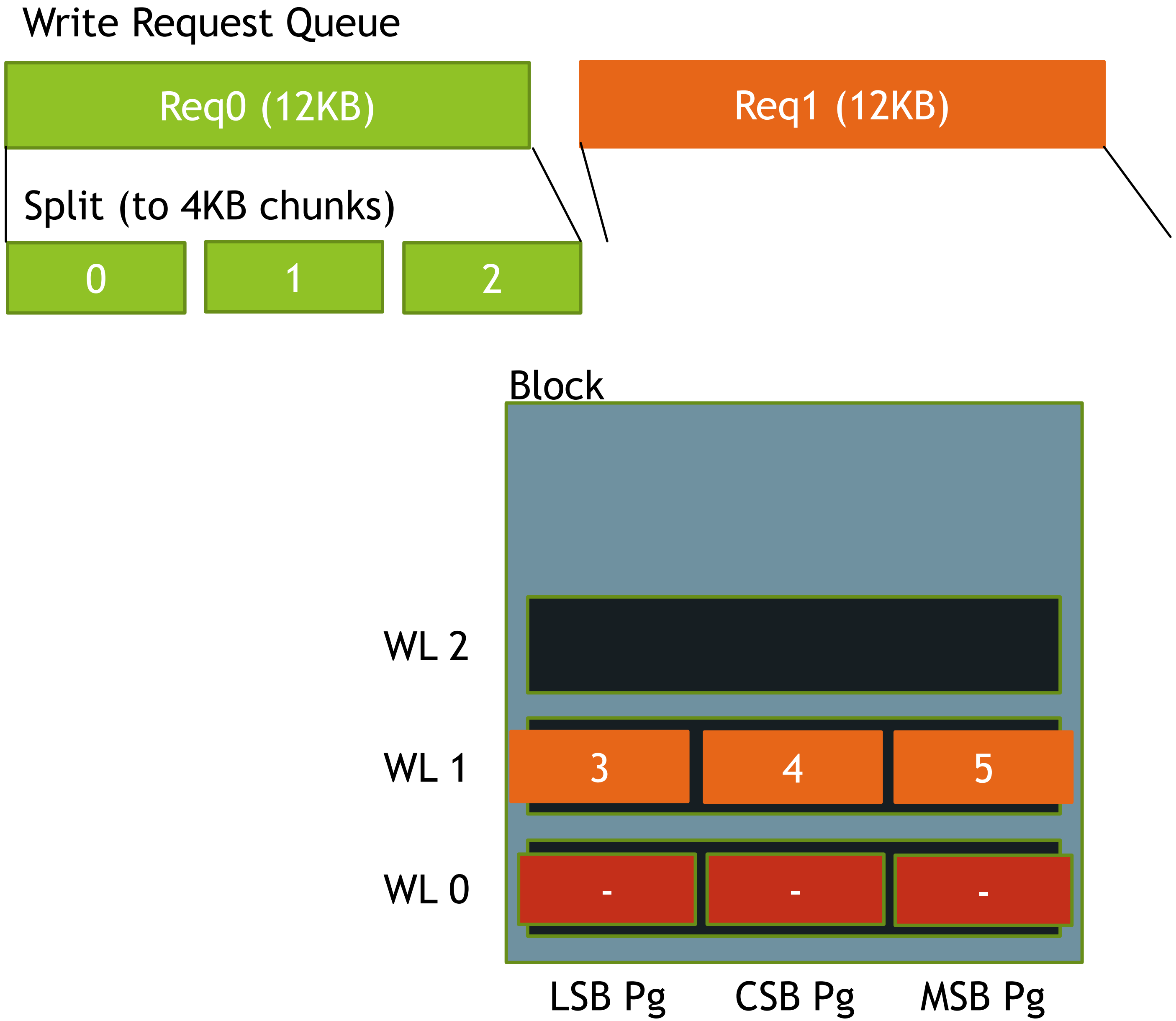
Write Request Queue



Scheduling of Writes

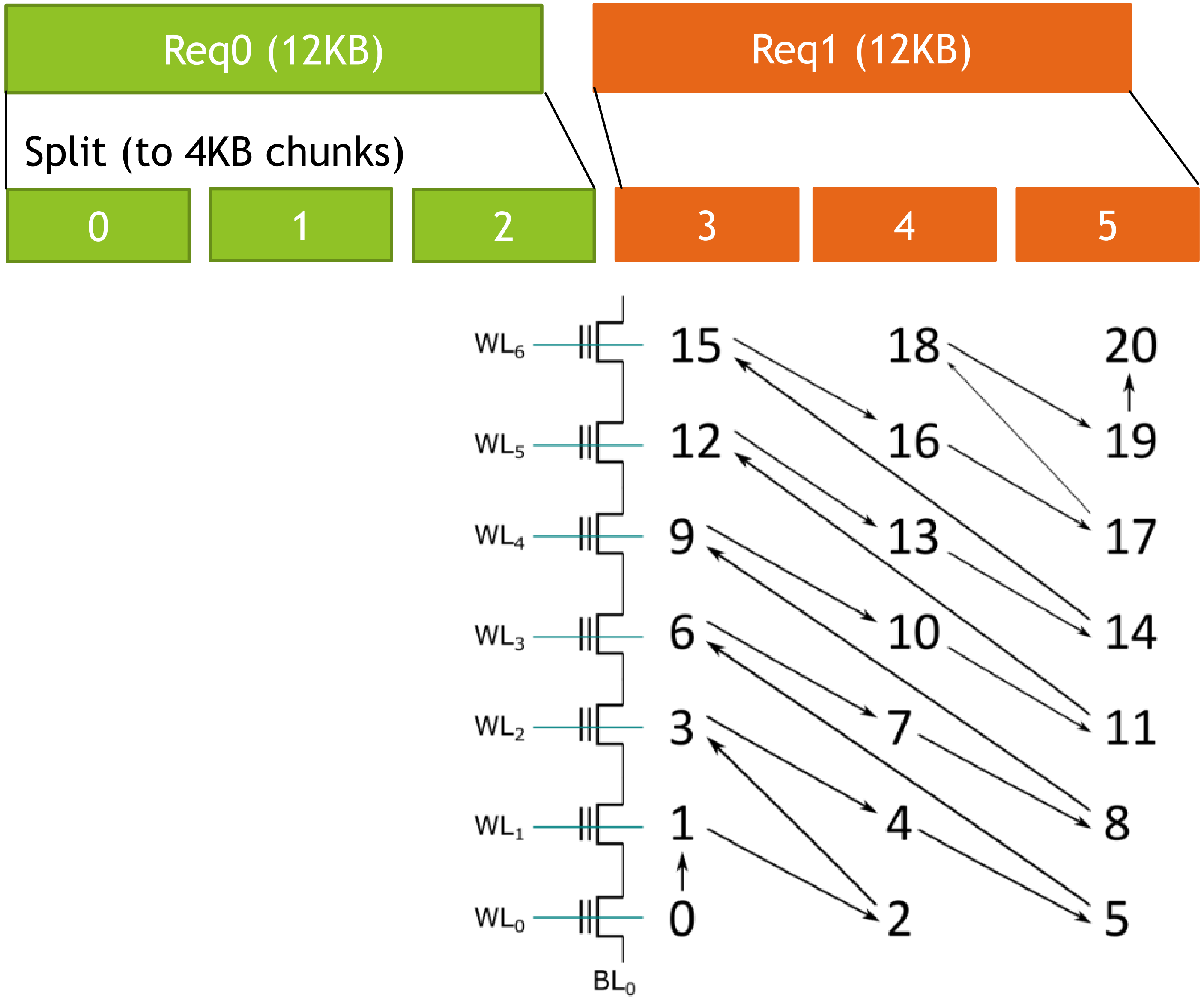


Scheduling of Writes



Scheduling of Writes

Write Request Queue



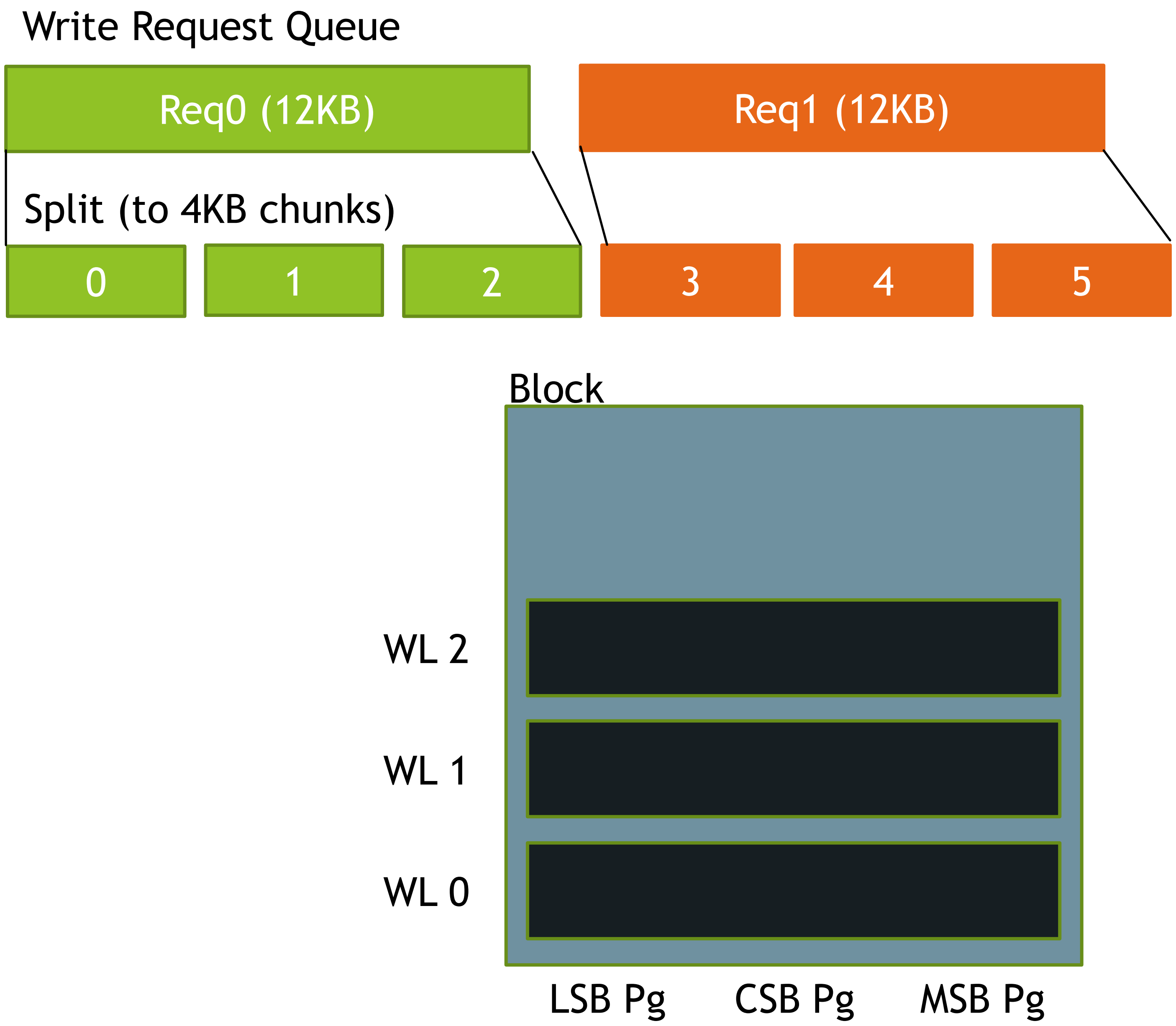
Req0 (12KB)

Req1 (12KB)

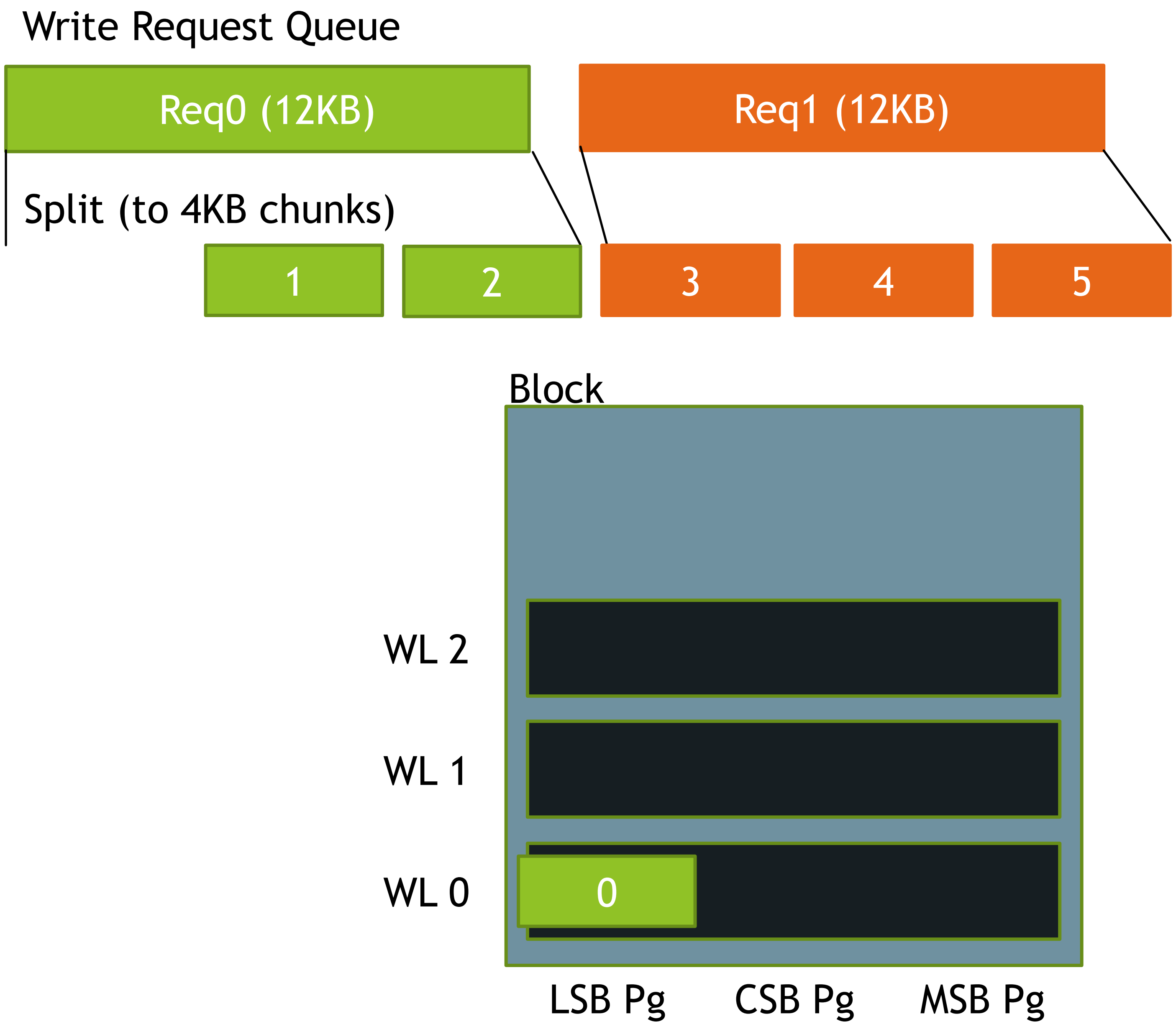
Split (to 4KB chunks)

0 1 2 3 4 5

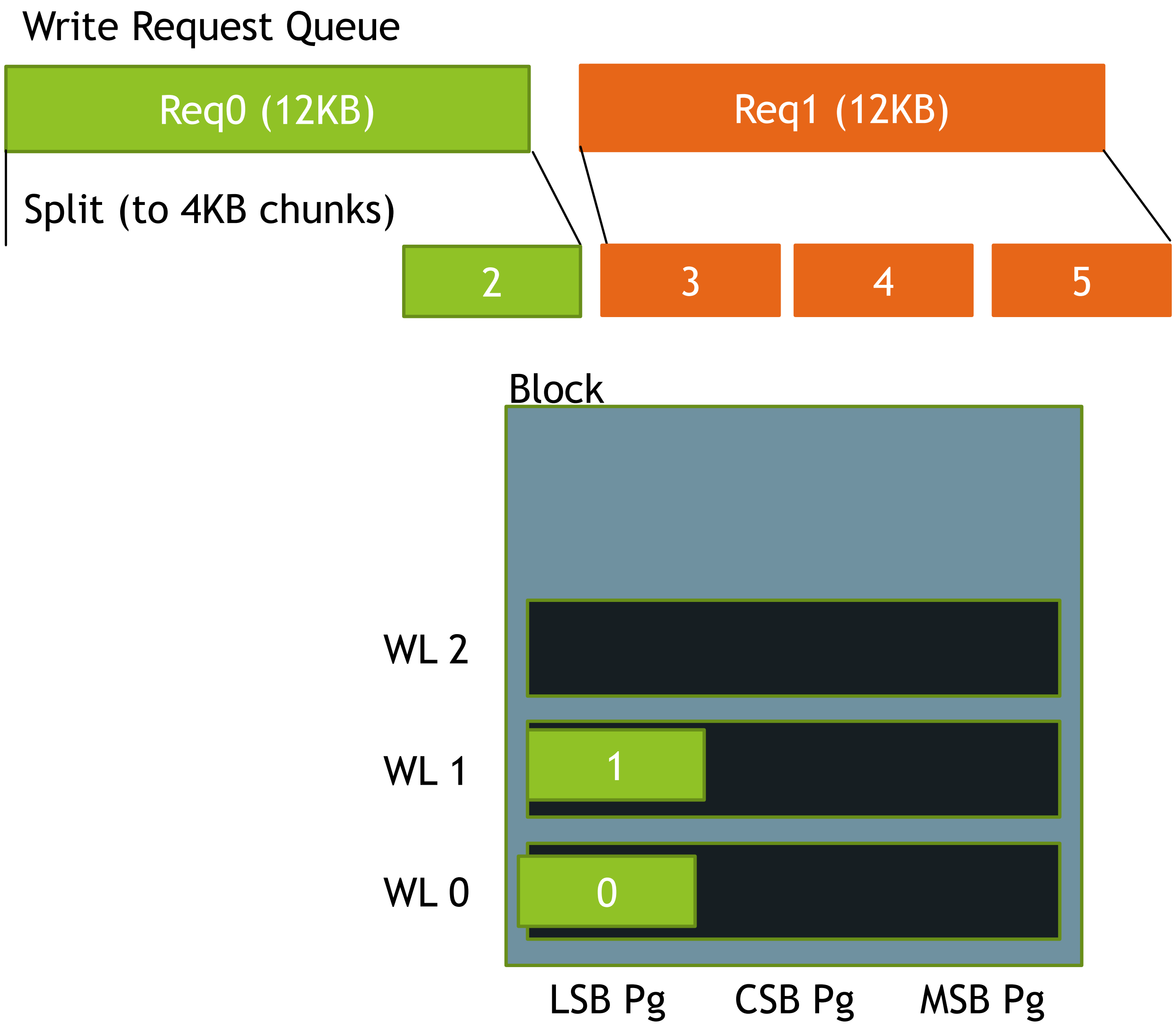
Scheduling of Writes



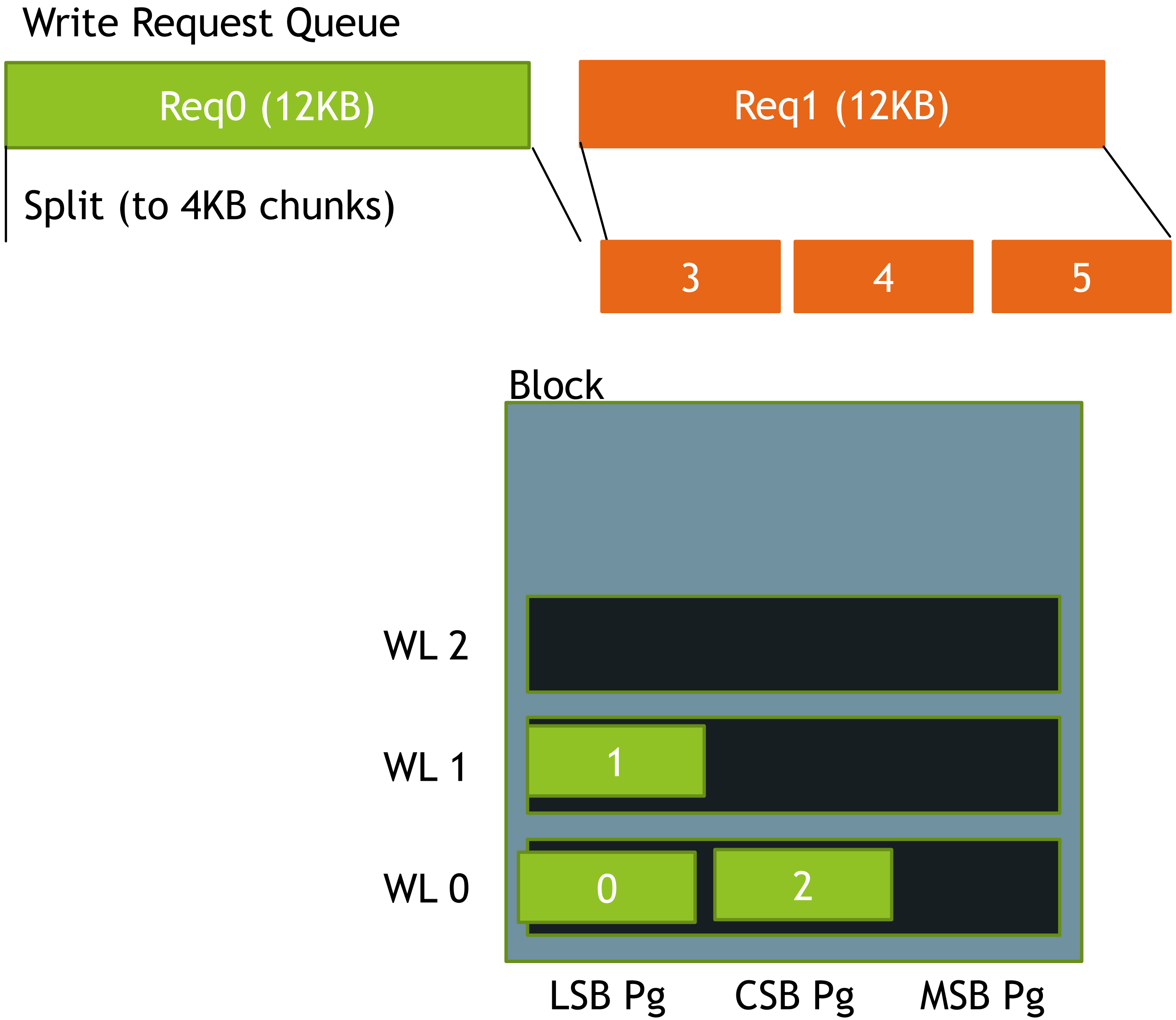
Scheduling of Writes



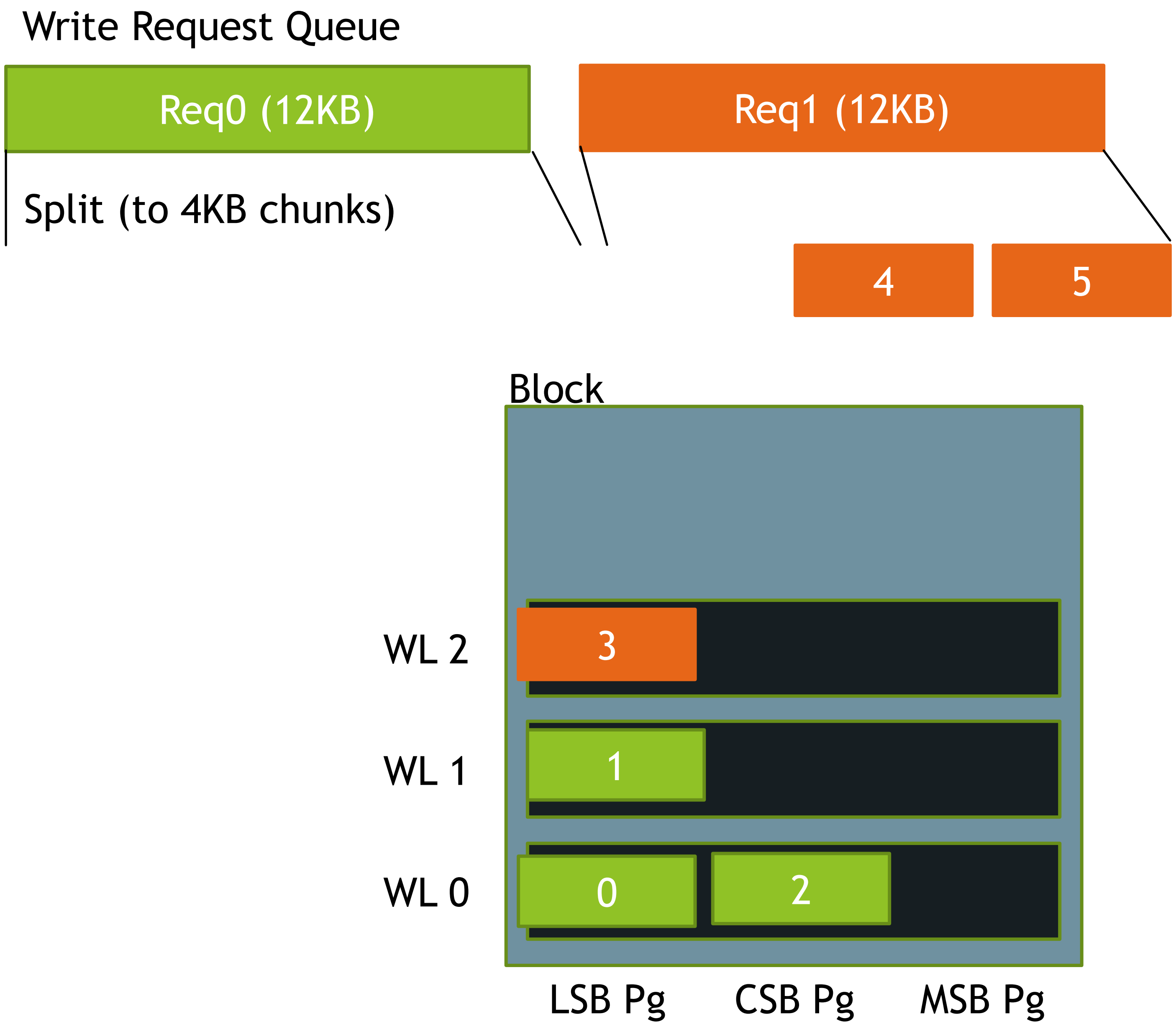
Scheduling of Writes



Scheduling of Writes

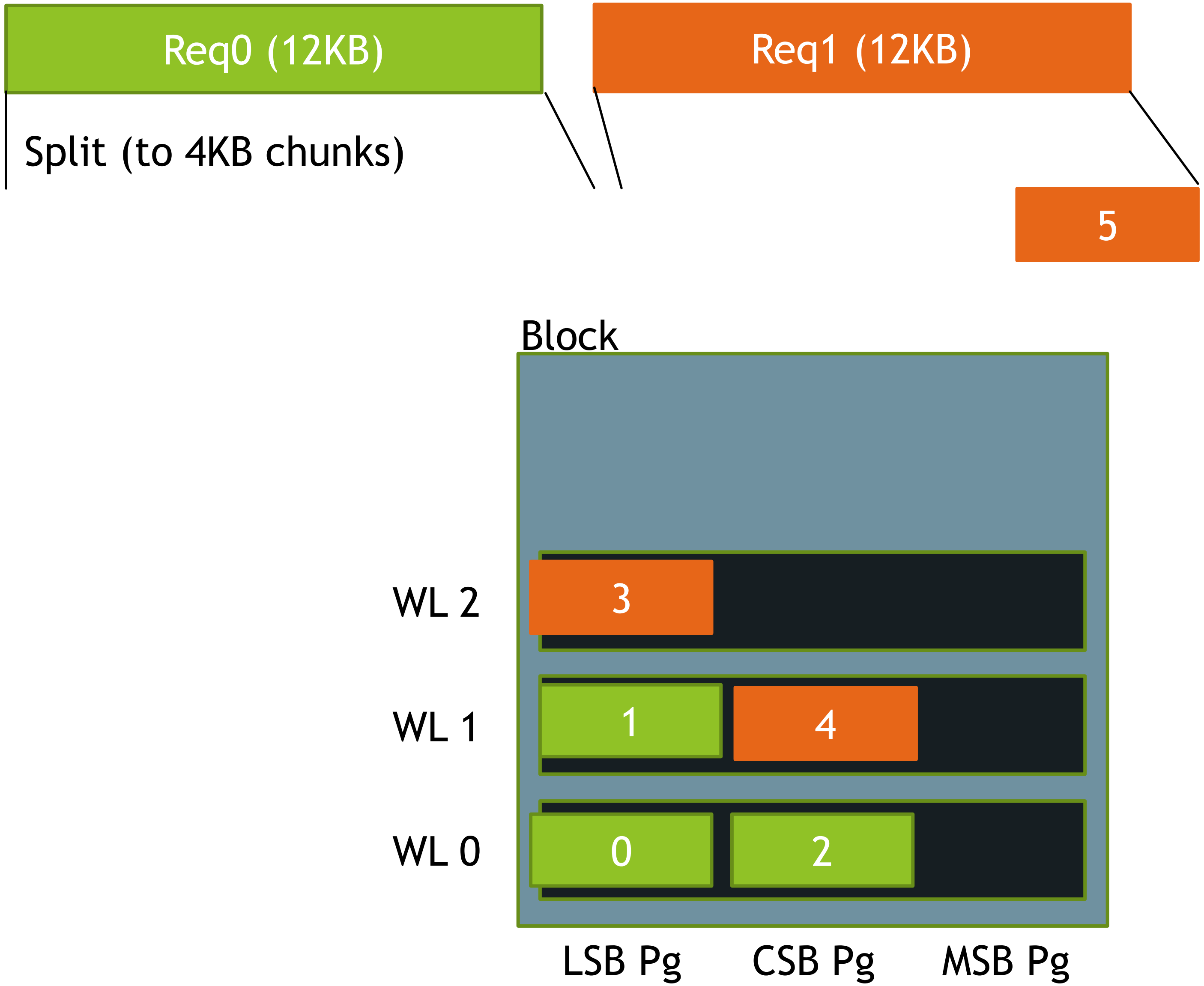


Scheduling of Writes

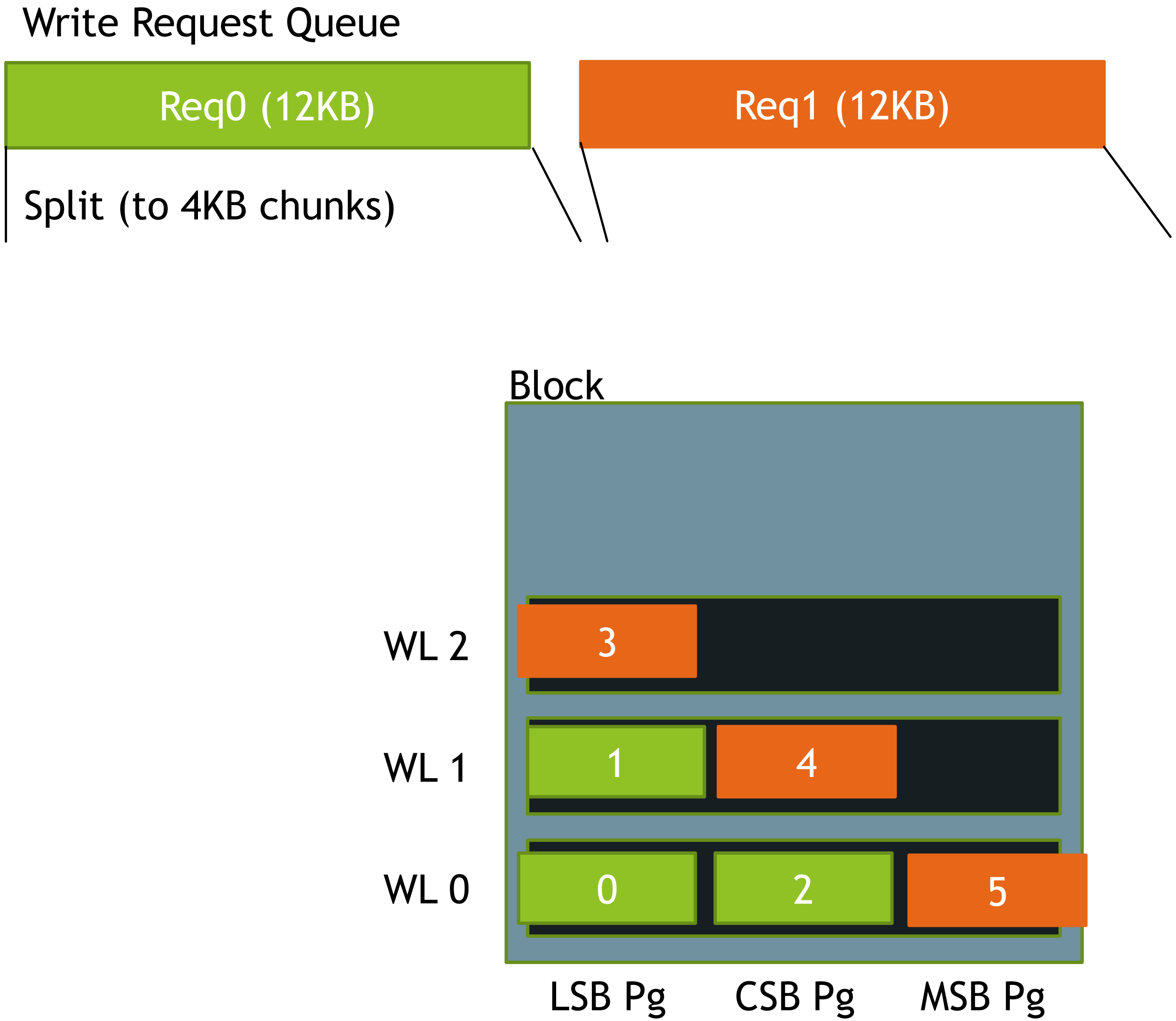


Scheduling of Writes

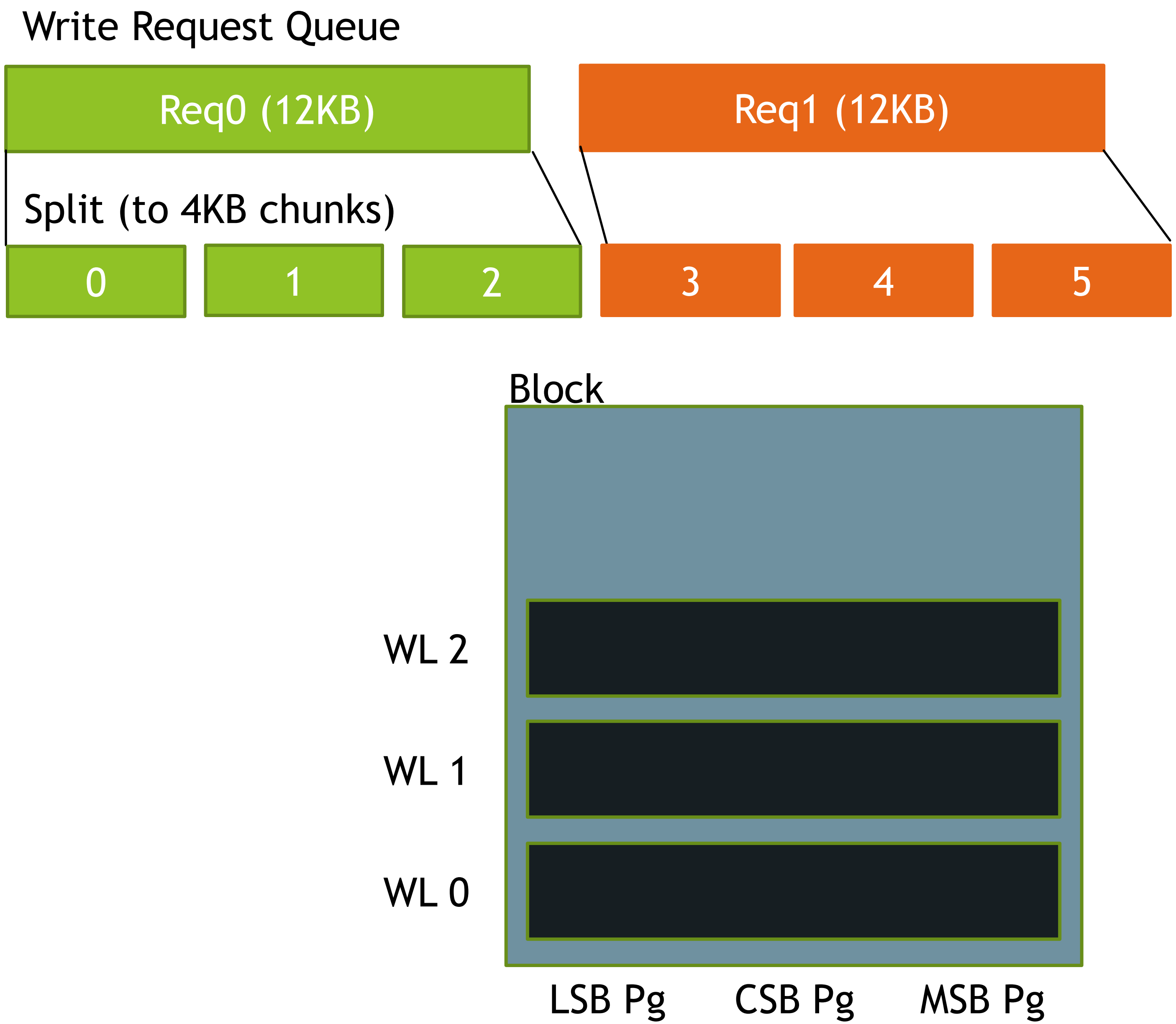
Write Request Queue



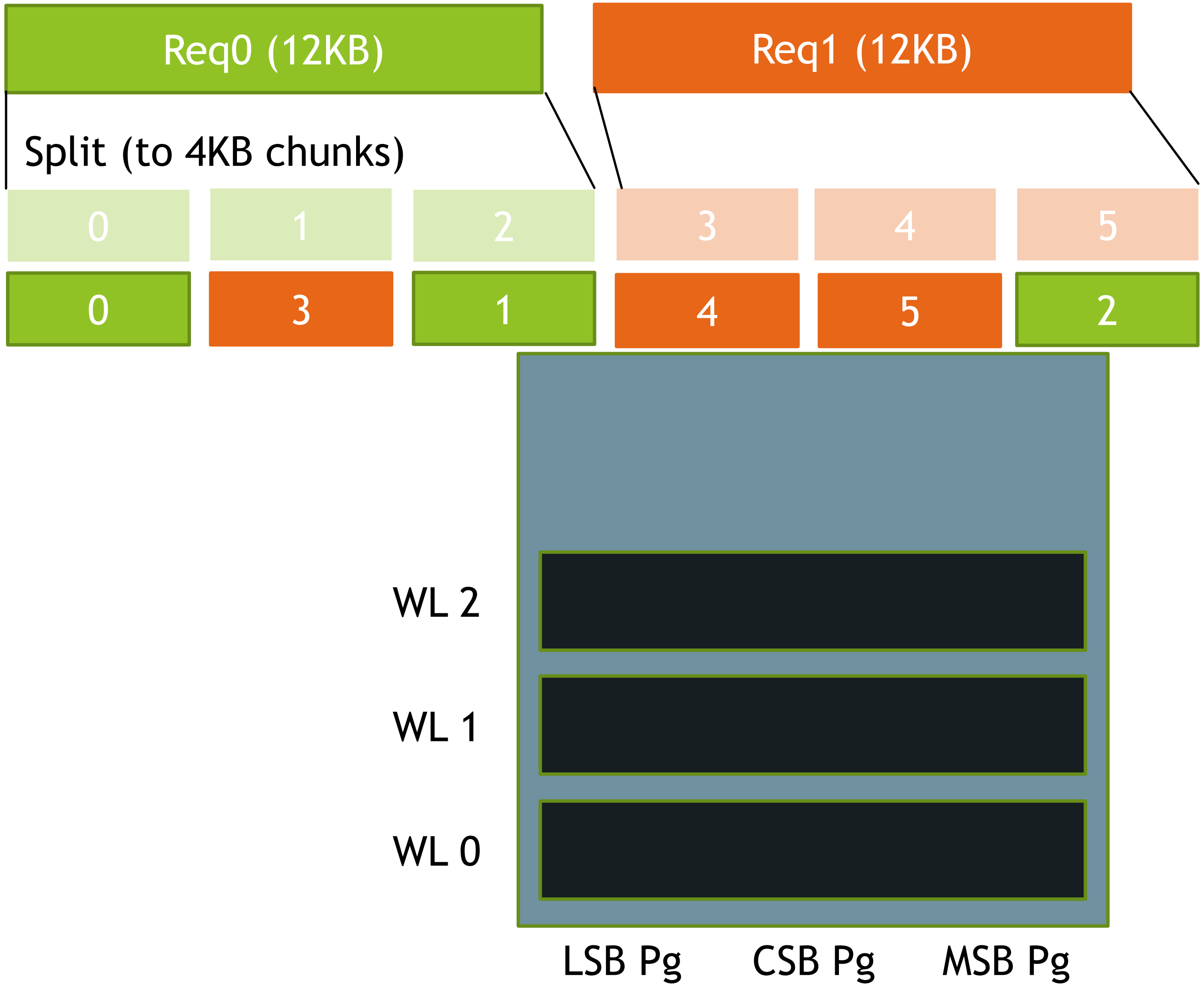
Scheduling of Writes



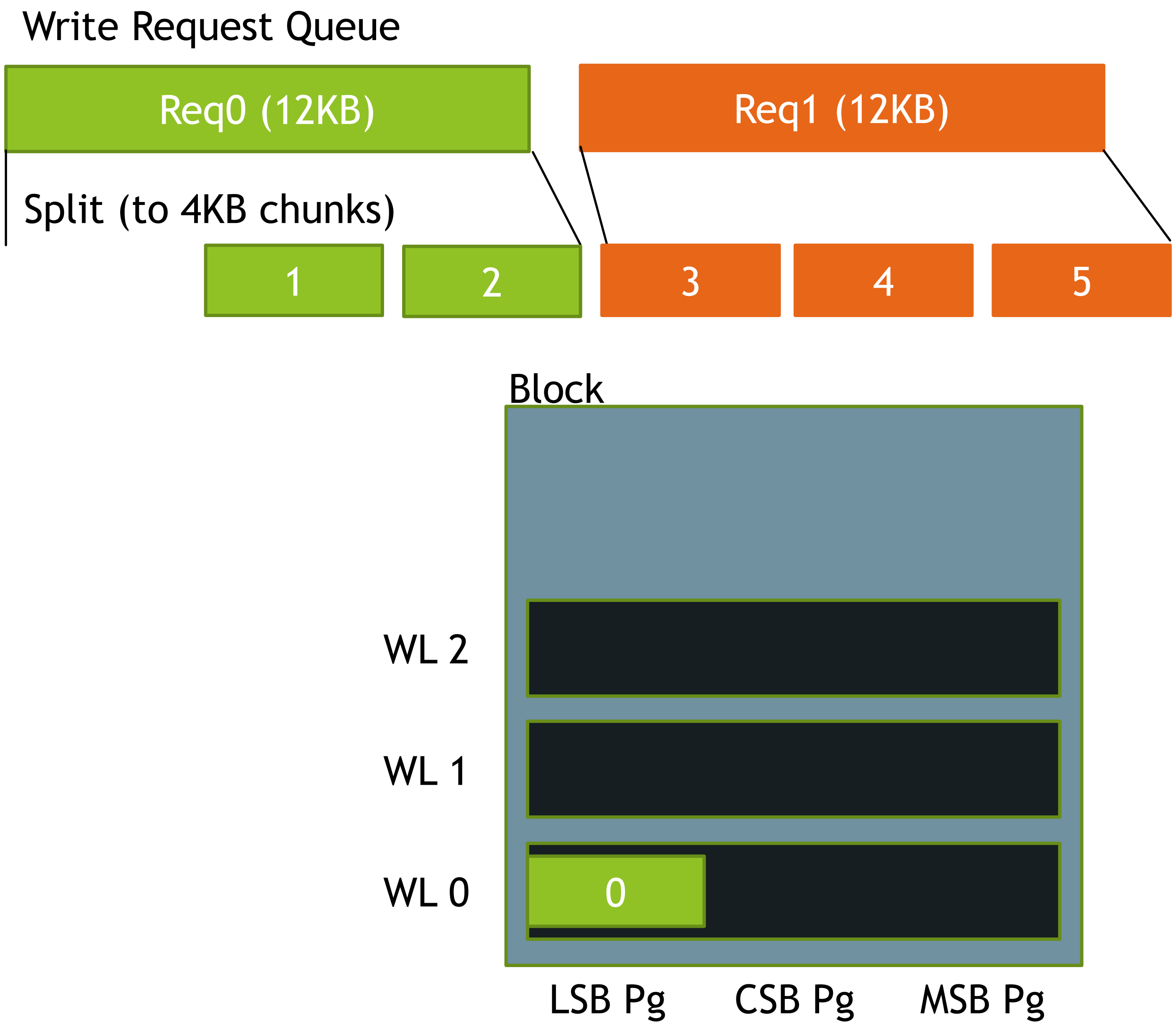
Scheduling of Writes



Scheduling of Writes

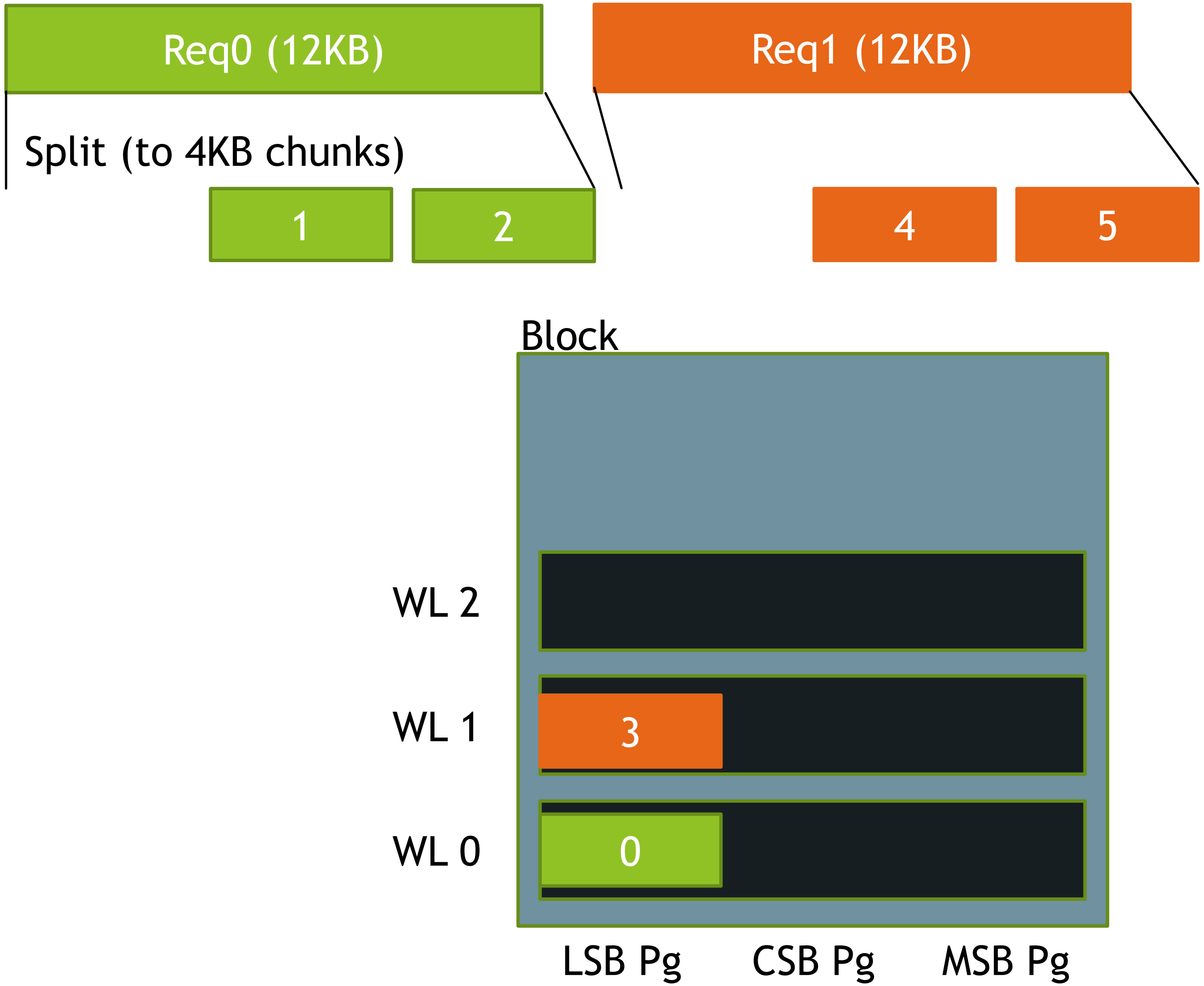


Scheduling of Writes



Scheduling of Writes

Write Request Queue



Req0 (12KB)

Req1 (12KB)

Split (to 4KB chunks)

1

2

4

5

Block

WL 2

WL 1

3

WL 0

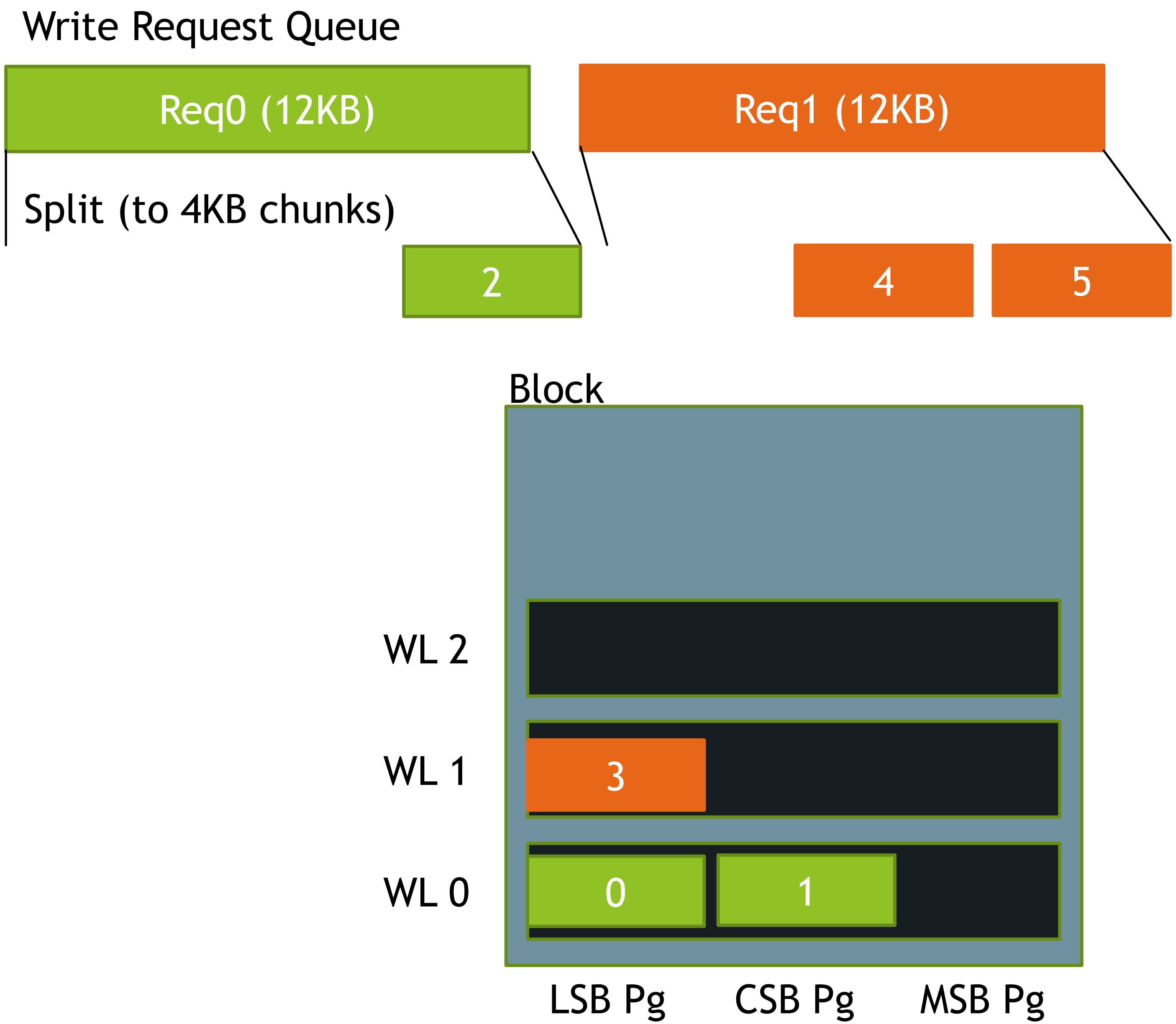
0

LSB Pg

CSB Pg

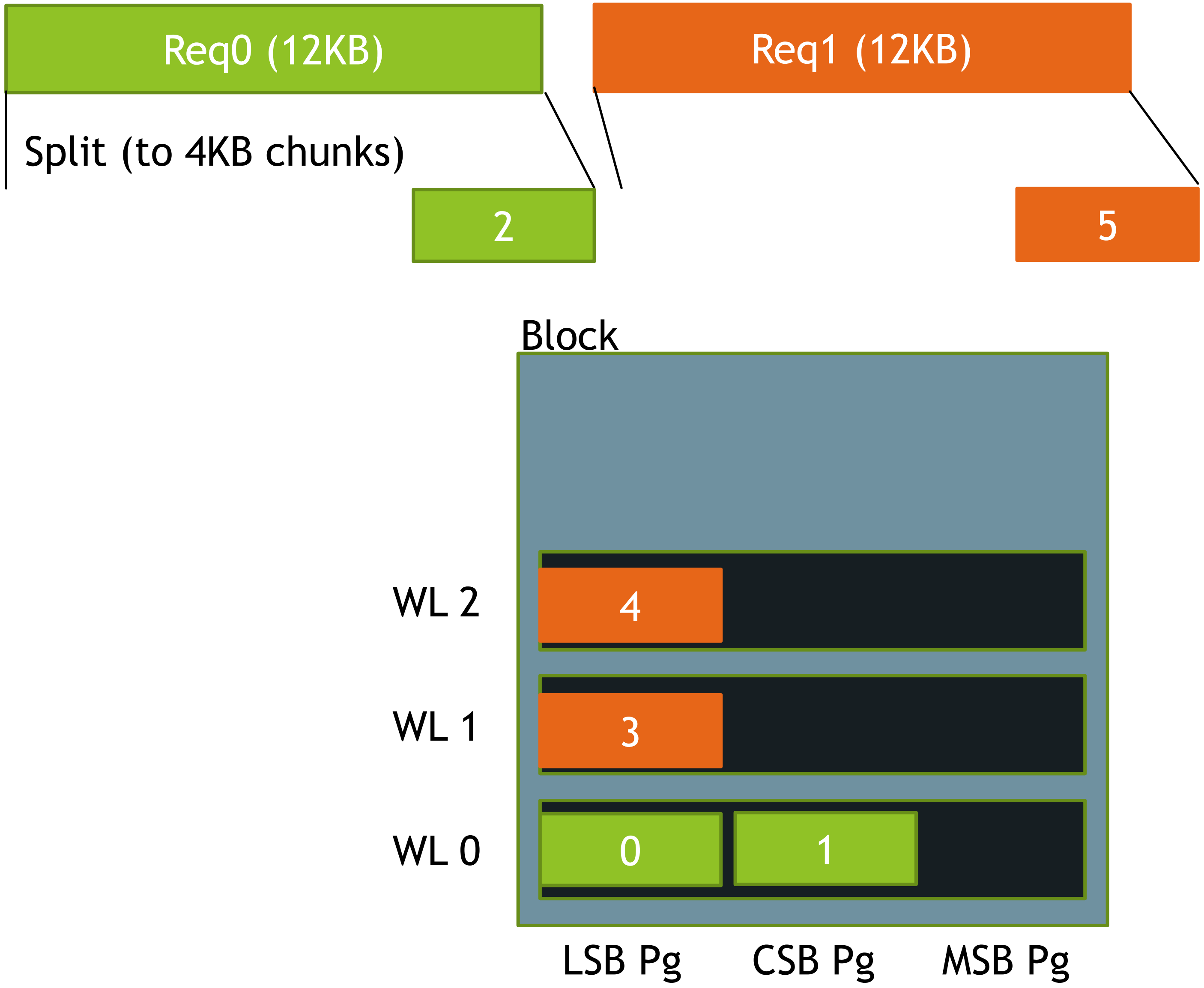
MSB Pg

Scheduling of Writes



Scheduling of Writes

Write Request Queue



Req0 (12KB)

Req1 (12KB)

Split (to 4KB chunks)

2

5

Block

WL 2

4

WL 1

3

WL 0

0

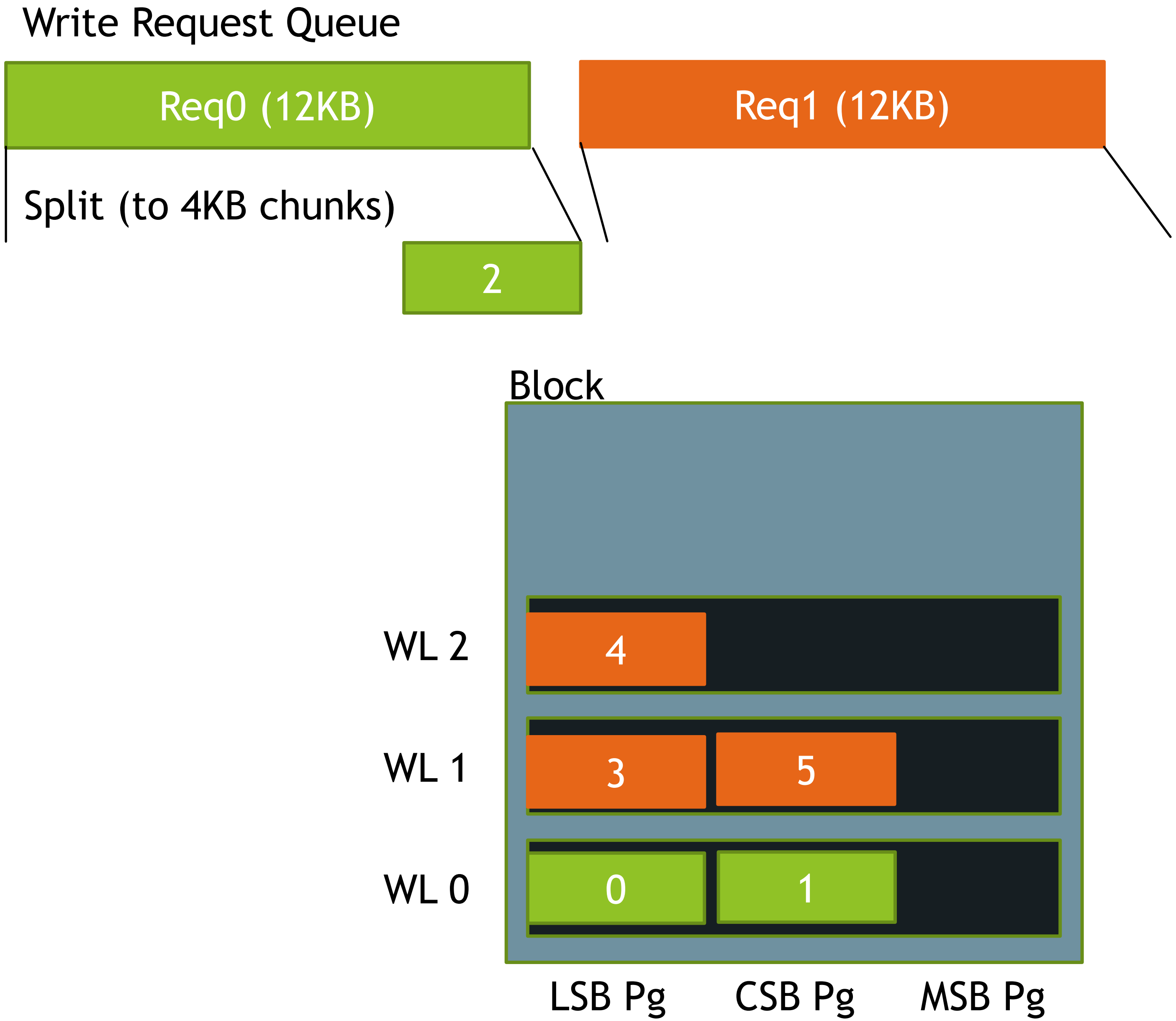
1

LSB Pg

CSB Pg

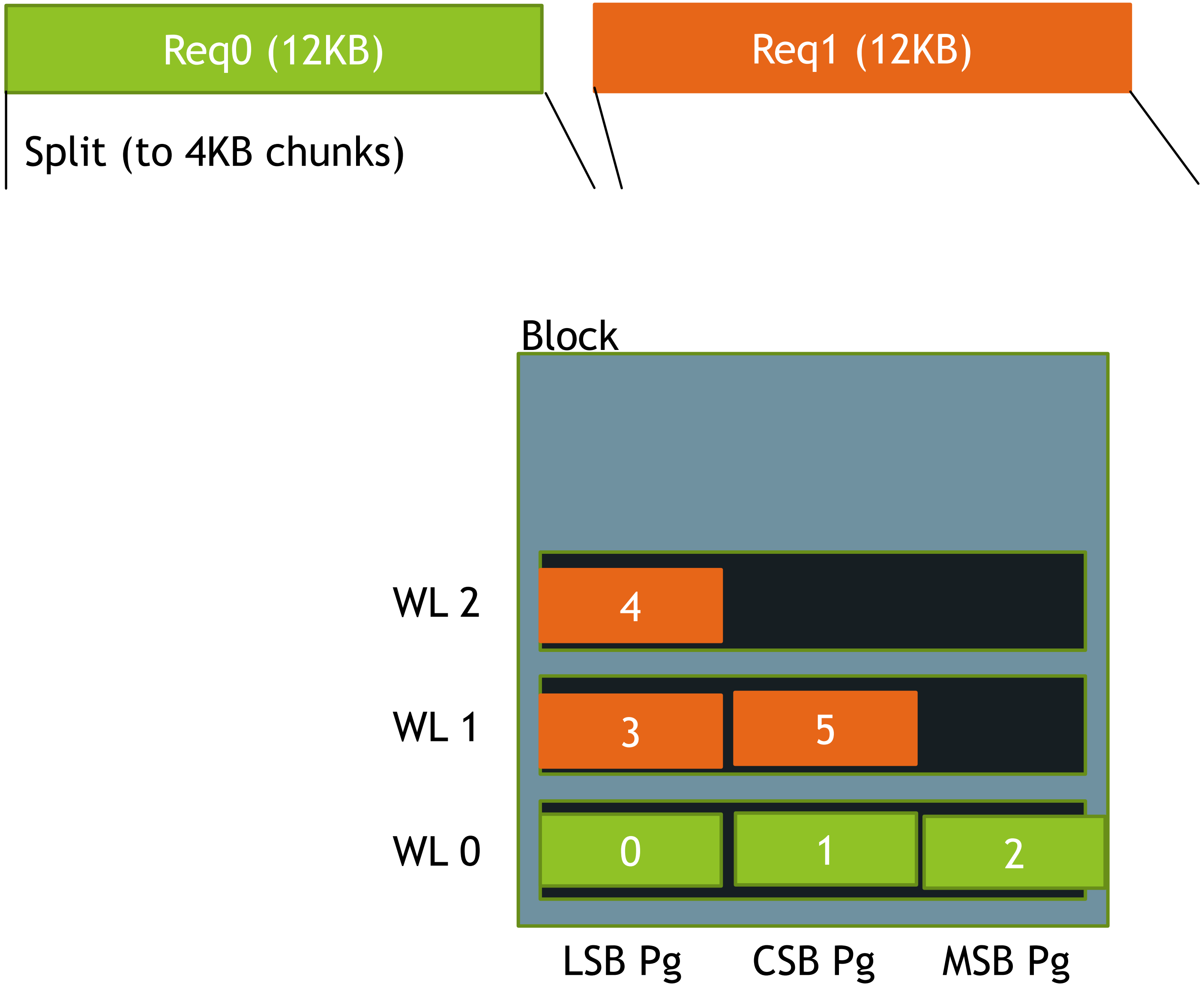
MSB Pg

Scheduling of Writes



Scheduling of Writes

Write Request Queue



Req0 (12KB)

Req1 (12KB)

Split (to 4KB chunks)

Block

WL 2

4

WL 1

3

5

WL 0

0

1

2

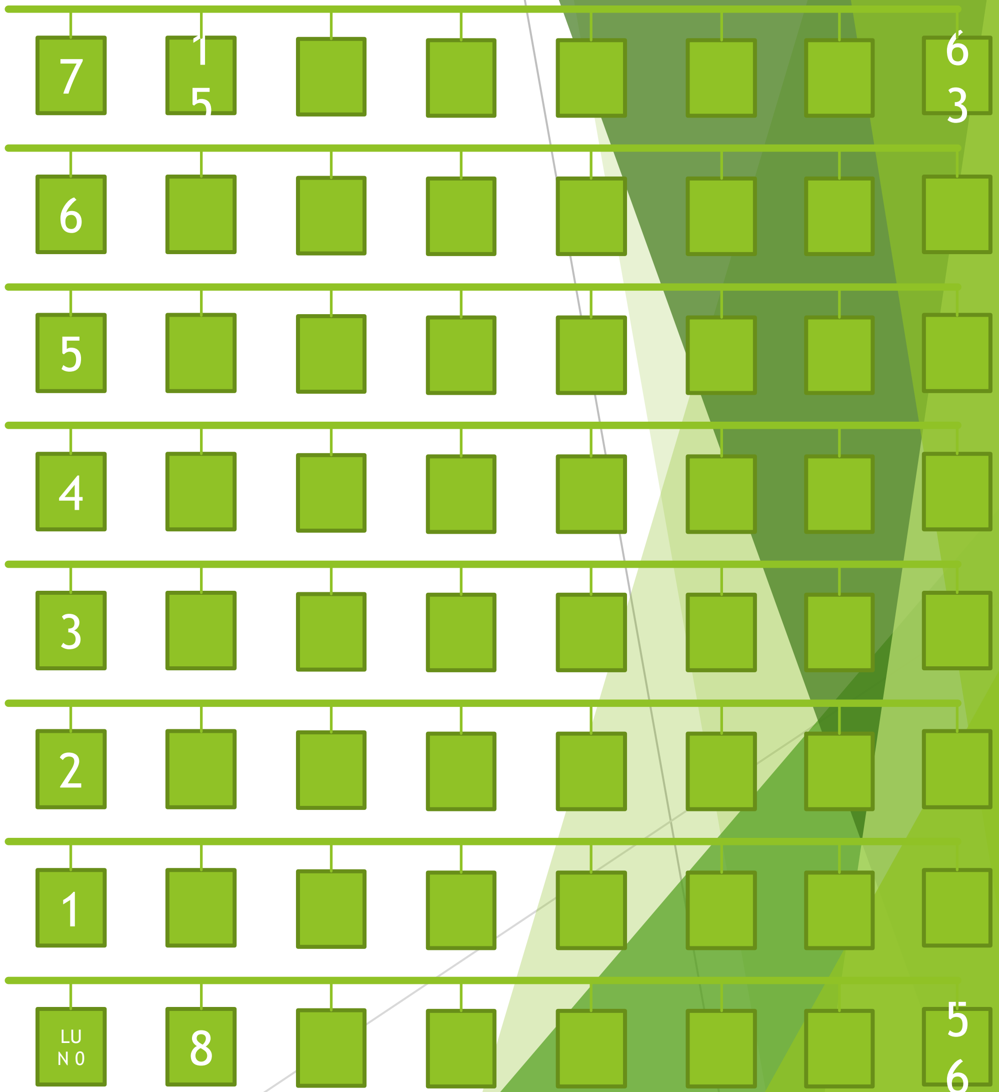
LSB Pg

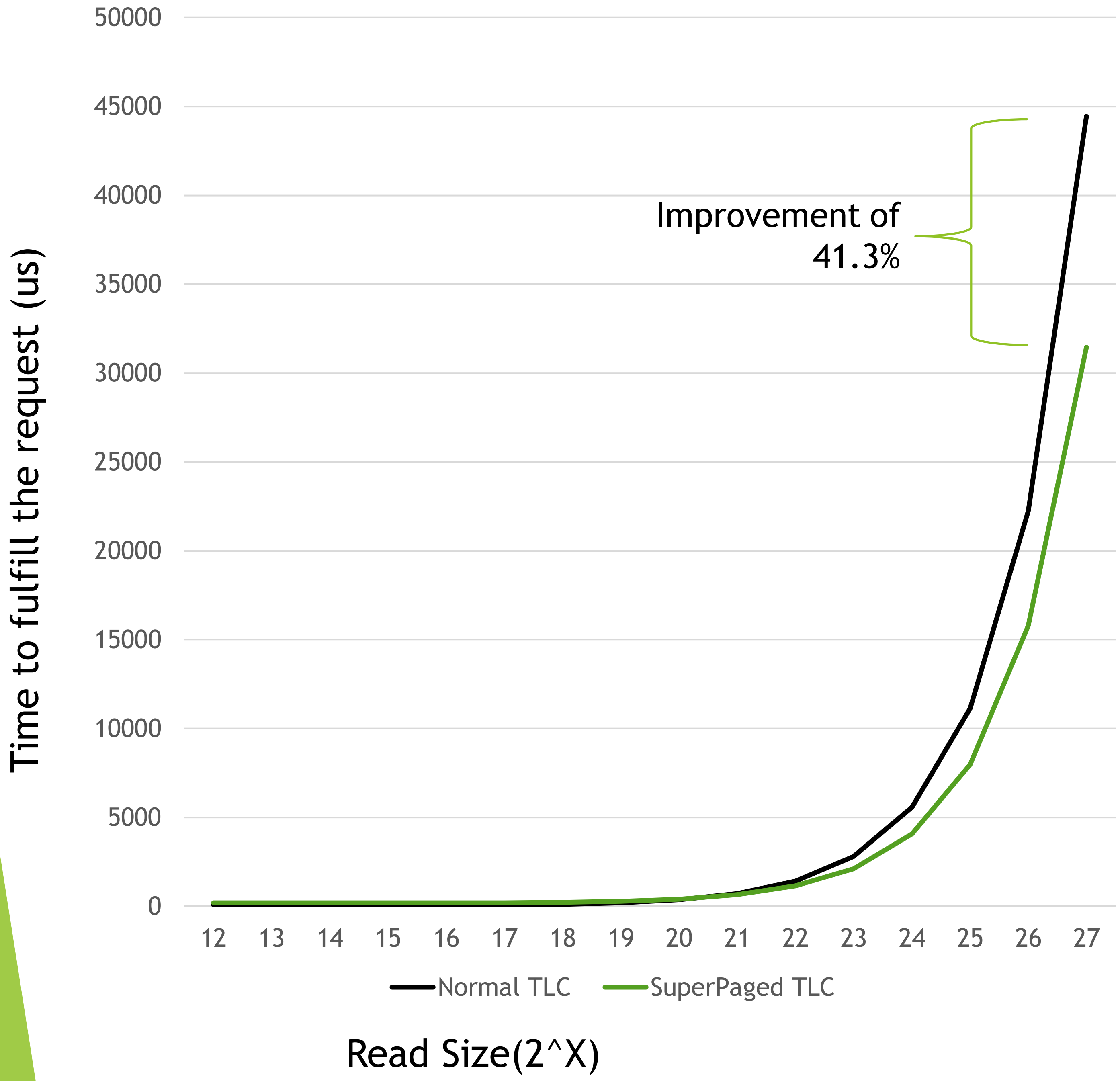
CSB Pg

MSB Pg

- It’s only beneficial to use melded pages when large amounts of data needs to be read.
- How large is large enough?

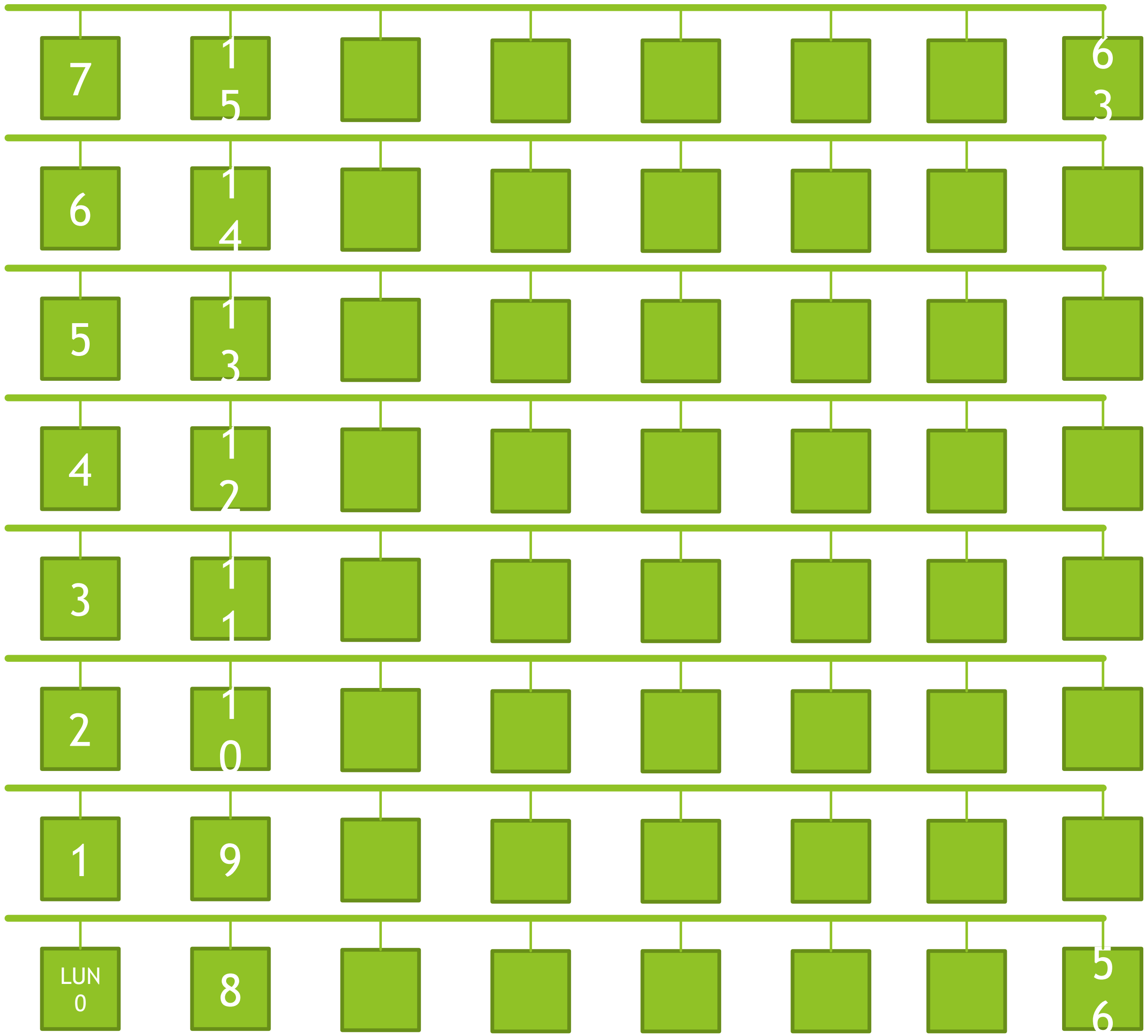
- Number of channels: 8
- Number of parallel units per channel: 8
- Total number if parallel units: 64
- Channel's operating frequency : 800 MT/s
- Page Size: 4KB





Improvement of 41.3%

	Normal TLC (us)	Melded TLC (us)
2^12	63	183
2^13	63	183
2^14	63	183
2^15	63	183
2^16	69	183
2^17	81	200
2^18	104	218
2^19	188	270
2^20	364	401
2^21	708	636
2^22	1406	1134
2^23	2791	2103
2^24	5572	4068
2^25	11124	7971
2^26	22236	15803
2^27	44452	31440



	Normal TLC (us)	Melded TLC (us)
2^12	63	183
2^13	63	183
2^14	63	183
2^15	63	183
2^16	69	183
2^17	81	200
2^18	104	218
2^19	188	270
2^20	364	401
2^21	708	636
2^22	1406	1134
2^23	2791	2103
2^24	5572	4068
2^25	11124	7971
2^26	22236	15803
2^27	44452	31440

- It’s only beneficial to use melded pages when large amounts of data needs to be read.
- Problem: Decision to use melded pages needs to be done in program phase.
- How does the scheduler know the read pattern during writes.

Directives (Hints)

- Host provides hints to the scheduler when submitting the write request.
- NVMe's Directives support (1.3 and above)
  - Provides an ability to exchange extra metadata in the headers of ordinary NVMe commands.
  - Proposal is to add a new directive that enables the application to declare the read patterns.

Generating Hints

- Host provides hints to the scheduler when submitting the write request.
- These hints can be explicitly provided by the developer or automatically generated by looking at the history.

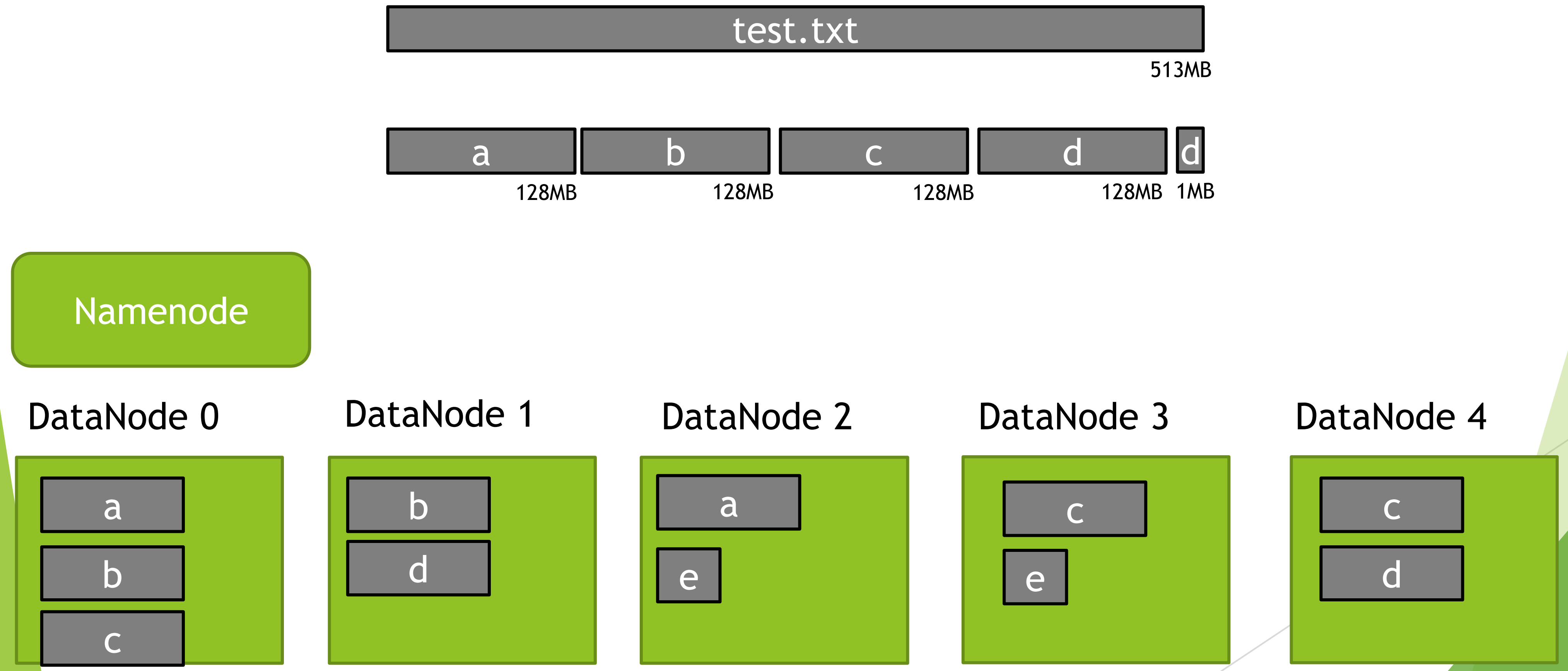
Hadoop Distributed File System

- **Hadoop** and **Spark** is an open-source cluster-computing framework.
- Large-scale data processing.
- Data itself is managed using HDFS.
  - HDFS is designed to store very large files across machines in a large cluster.

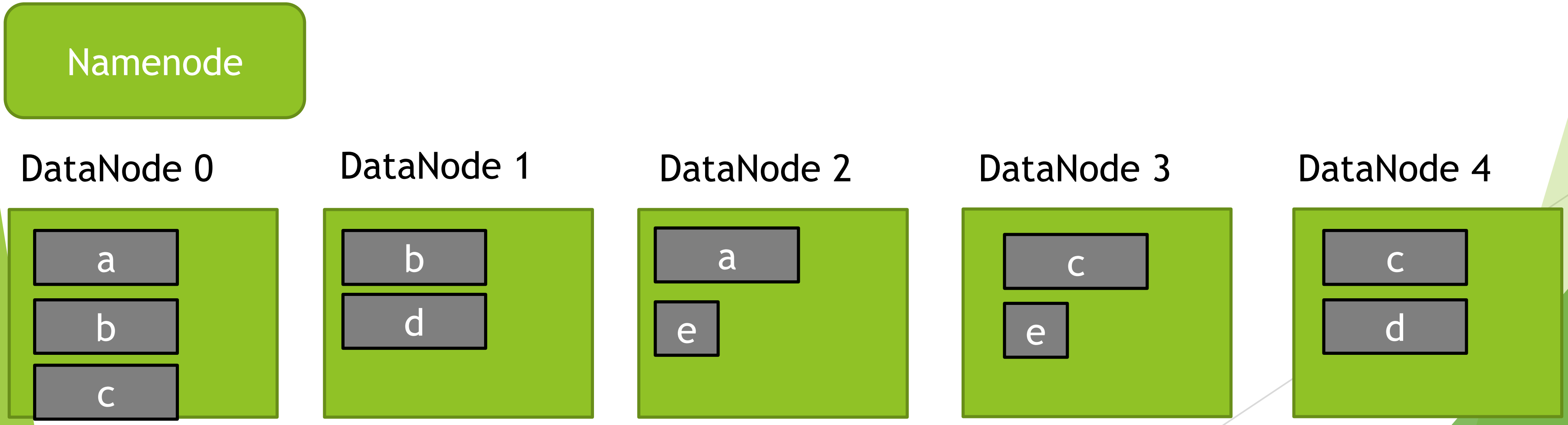
Hadoop Distributed File System

- **NameNodes**
  - HDFS cluster consists of a single NameNode.
  - Manages metadata
  - Maintains mapping of blocks to DataNodes
- **DataNodes**
  - Usually one per node in the cluster.
  - Stores blocks of data.

- When you store a file in HDFS, the system breaks it down into a set of individual blocks and stores these blocks in various data nodes in the Hadoop cluster.
- In HDFS, block size, by default, is 128 MB.



- To read a file, HDFS client first asks the NameNode for the list of DataNodes that host replicas of the blocks of the file.
- The client contacts a DataNode directly and requests the transfer of the desired block.
- Why large block size?



- To read a file, HDFS client first asks the NameNode for the list of DataNodes that host replicas of the blocks of the file.
- The client contacts a DataNode directly and requests the transfer of the desired block.
- Why large block size?
  - Assume we need to manage 1TB of data.
  - Number of entries in namenode (with 4K block size): 268,453,456
  - Number of entries in namenode (with 128M block Size): 8,192

		400MT/s (8 bits/transfer)		800MT/s (8 bits/transfer)		1600MT/s (8 bits/transfer)		1600MT/s (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
2KB (6KB)	Throughput (MBPS)	1440	2038	1490	2141	1516	2196	1530	2225
	% improvement	41.5%		43.6%		44.8%		45.4%	

		400MT/s (8 bits/transfer)		800MT/s (8 bits/transfer)		1600MT/s (8 bits/transfer)		1600MT/s (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
4KB (12KB)	Throughput (MBPS)	2466	2691	2879	4071	2980	4279	3033	4391
	% improvement	9.1%		41.3%		43.5%		44.7%	

		400MT/s (8 bits/transfer)		800MT/s (8 bits/transfer)		1600MT/s (8 bits/transfer)		1600MT/s (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
8KB (24KB)	Throughput (MBPS)	2697	2691	4930	5364	5756	8100	5960	8512
	% improvement	-		8.8%		40.7%		42.8%	

		400MT/s (8 bits/transfer)		800MT/s (8 bits/transfer)		1600MT/s (8 bits/transfer)		1600MT/s (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
16KB (48KB)	Throughput (MBPS)	2698	2688	5390	5357	9849	10641	11507	16060
	% improvement	-		-		8.0%		39.5%	

Read throughputs of SSD (8 channels; 8 parallel units per channel)}

		400MT/s (8 bits/transfer)		**800MT/s** (8 bits/transfer)		**1600MT/s** (8 bits/transfer)		**1600MT/s** (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
2KB (6KB)	Throughput (MBPS)	1440	2040	1490	2141	1516	2196	1530	2225
	% improvement	41.6%		43.6%		44.8%		45.4%	

		400MT/s (8 bits/transfer)		**800MT/s** (8 bits/transfer)		**1600MT/s** (8 bits/transfer)		**1600MT/s** (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
4KB (12KB)	Throughput (MBPS)	2699	3721	2880	4078	2981	4282	3033	4393
	% improvement	37.8%		41.5%		43.6%		44.8%	

		400MT/s (8 bits/transfer)		**800MT/s** (8 bits/transfer)		**1600MT/s** (8 bits/transfer)		**1600MT/s** (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
8KB (24KB)	Throughput (MBPS)	4624	5357	5398	7401	5762	8109	5963	8516
	% improvement	15.8%		37.1%		40.7%		42.8%	

		400MT/s (8 bits/transfer)		**800MT/s** (8 bits/transfer)		**1600MT/s** (8 bits/transfer)		**1600MT/s** (16 bits/transfer)	
Page Size		Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC	Normal TLC	Melded TLC
16KB (48KB)	Throughput (MBPS)	5390	5357	9241	10641	10794	14715	11531	16166
	% improvement	-		15.1%		36.3%		40.1%	

Read throughputs of SSD (16 channels; 4 parallel units per channel)}

Thank You

- Contact information of authors:
  - arpith@iisc.ac.in
  - gopi@iisc.ac.in