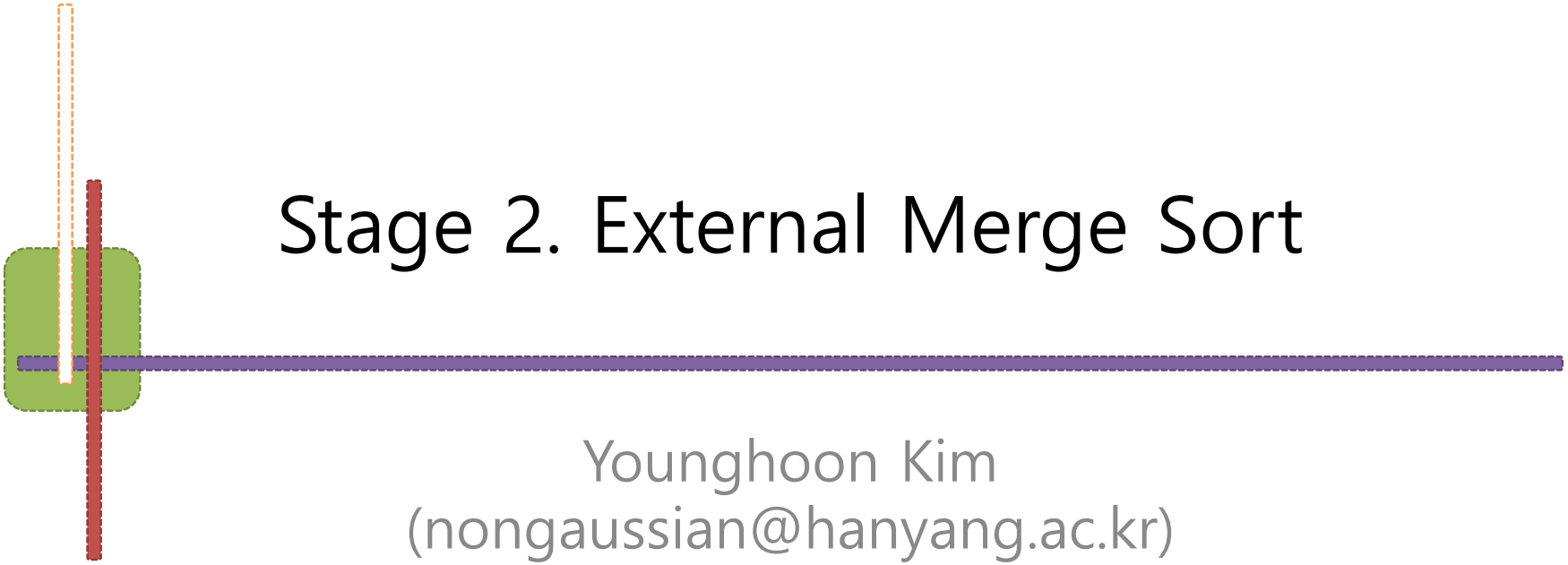


Stage 2. External Merge Sort

Younghoon Kim
(nongaussian@hanyang.ac.kr)

Goal

- Given
  - A file
    - Containing a list of triples with 3 integers (e.g., <5, 1, 2>)
    - For triples, use `org.apache.commons.lang3.tuple.MutableTriple`
- Return
  - A file
    - A list of triples sorted in the ascending order by using external merge sort
  - Sorting criteria
    - Primarily, sort by the first value
    - With tuples with an identical first value, use the second value
    - With tuples with identical first and second values, use the third value
  - Example

```
(4,8,4)(4,5,4)(7,9,6)(0,6,5)(6,0,3)(0,5,3)(3,1,7)(5,4,9)(4,6,6)(9,1,1)
```

```
(0,5,3)(0,6,5)(3,1,7)(4,5,4)(4,6,6)(4,8,4)(5,4,9)(6,0,3)(7,9,6)(9,1,1)
```

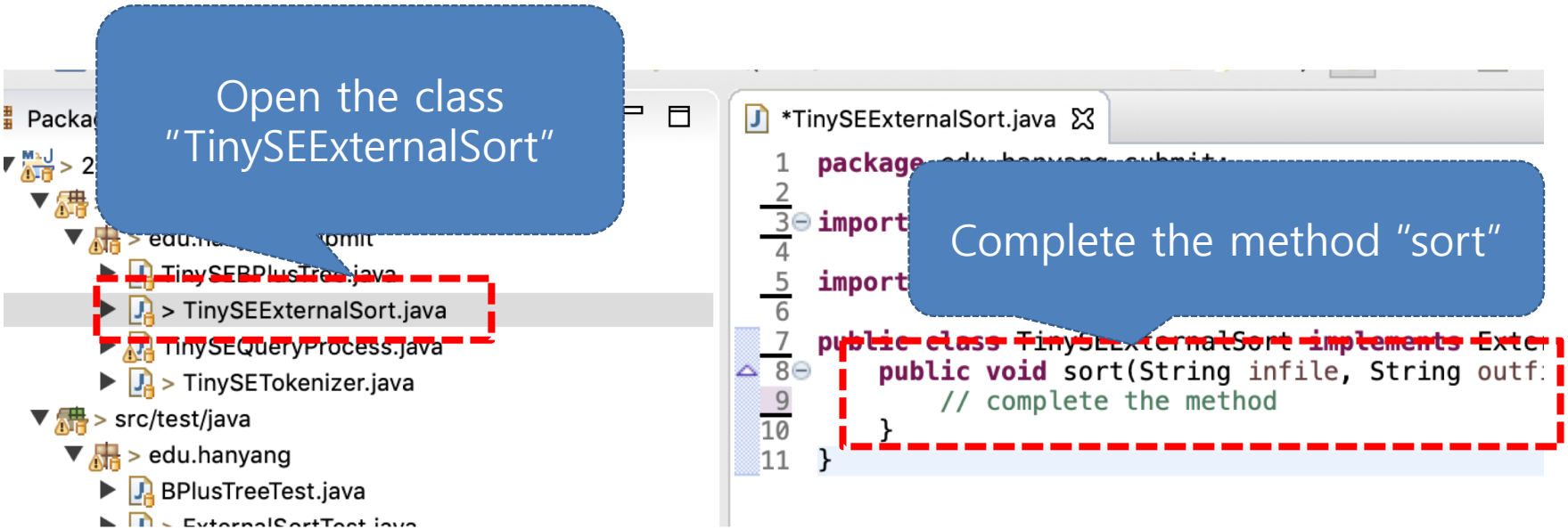
Code Template

- TinySE-submit
  - contains
    - Template codes (`edu.hanyang.submit.TinySEExternalSort.java`)
    - JUnit test codes
  - Depends on
    - TinySE framework (`<github>/nongaussian/tinyse`) ← to be updated on every stage
- TinySE
  - Includes
    - Interface files (e.g., ExternalSort.java)
    - Indexer and query processer codes which will complete a search engine by connecting your submissions

Complete Interface in TinySE-submit

- Step 1. Write your codes
  - Implement the method "sort" in the class "TinySEExternalSort"
- Step 2. Set @Ignore annotation
  - Delete comment test 2
  - Comment-out test 1, 3, 4
- Step 3. Test and build your codes
  - Run "mvn test" & "mvn package"
- Step 4. Submit your module
  - Commit & push into your TinySE-submit fork

Step 1. Write your codes



To Use Code Template

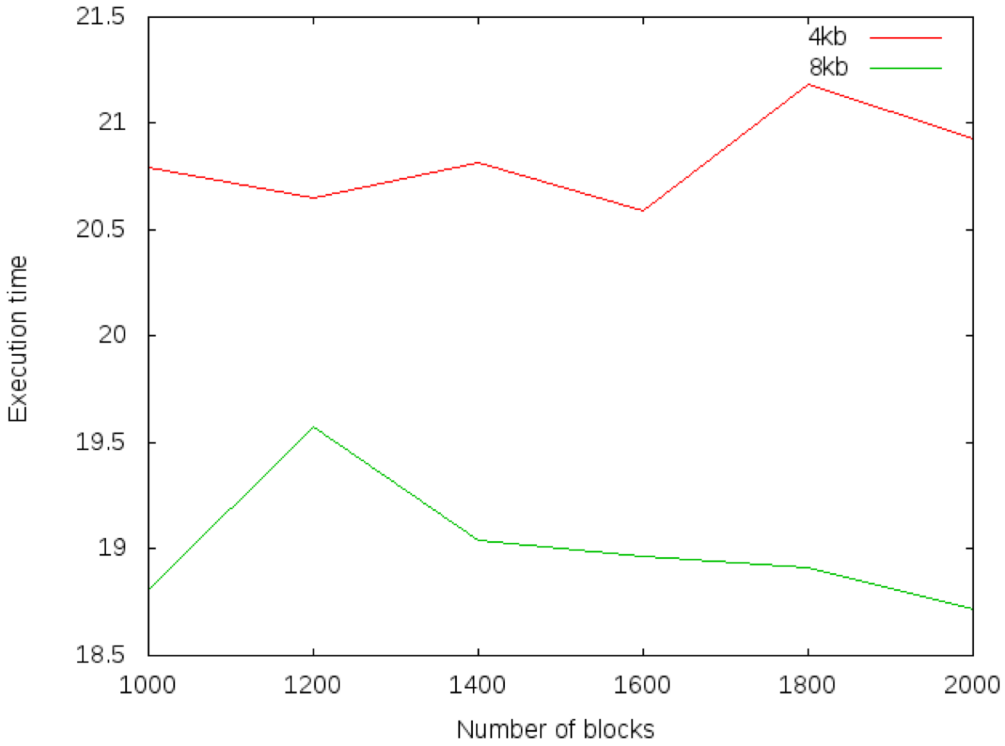
- Complete *edu.hanyang.submit.TinySEExternalSort*

```
package edu.hanyang.indexer;


public interface ExternalSort {
        public void sort(String infile,  /* input file path */
                         String outfile, /* output file path */
                         String tmpdir,  /* temporary directory
                                            for creating intermediate
                                            runs */
                         int blocksize,  /* 4096 or 8192 bytes */
                         int nblocks);   /* available memory size /
                                            blocksize */
}
```

An Example of Evaluation

- Test setting
  - Datasize: 100Mb
  - Heapsize: 16Mb
  - Blocksize: 4kb, 8kb
  - # of blocks: 1000, 1200, …, 2000



Read and Write Buffers

- To implement an external sort,
  - You may need read a given size of blocks sequentially for each run
  - How do we implement it in Java?
  - → Use BufferedReader

Utility Class

- DiskIO.class
  - edu.hanyang.utils.DiskIO;

Method Summary

All Methods | Static Methods | Concrete Methods

Modifier and Type	Method and Description
static void	**append_arr**(java.io.DataOutputStream out, java.util.List<org.javatuples.Triplet<java.lang.Integer,java.lang.Integer,java.lang.Integer>> arr, int nelements) Write the data which in 'arr' from zero to 'nelements', to file.
static java.io.DataInputStream	**open_input_run**(java.lang.String filepath) Create and return DataInputStream instance.
static java.io.DataOutputStream	**open_output_run**(java.lang.String filepath) Create and return DataOutputStream instance.
static int	**read_array**(java.io.DataInputStream in, int offset, int nelements, java.util.ArrayList<org.javatuples.Triplet<java.lang.Integer,java.lang.Integer,java.lang.Integer>> arr) Read Triplet data from DataInputStream and insert into given ArrayList.
static void	**sort_arr**(java.util.List<org.javatuples.Triplet<java.lang.Integer,java.lang.Integer,java.lang.Integer>> arr, int nelements) Sort the Triplet which in given ArrayList from zero to 'nelements'.

Methods inherited from class java.lang.Object

equals, getClass, hashCode, notify, notifyAll, toString, wait, wait, wait

Step 2. Set @Ignore annotation

- Delete comment test 2

ExternalSortTest.java

```
1  package edu.hanyang;
2
3  import static org.junit.Assert.*;
16
17 //@Ignore("Delete this line to unit test stage 2")
18 public class ExternalSortTest {
19     @Before
20     public void init() {
21         clean("./tmp");
22         File resultFile = new File("./sorted.data");
23         if(resultFile.exists()) {
24             resultFile.delete();
25         }
26     }
27
```

- Comment-out test 1, 3, 4

*TokenizerTest.java | ExternalSortTest.java

```
1  package edu.hanyang;
2
3  import static org.junit.Assert.assertTrue;
15
16 @Ignore("Delete this line to unit test stage 1")
17 public class TokenizerTest {
18     static List<String[]> results;
19     static List<String> testSentences;
20
21     @BeforeClass
22     public static void init() {
23         results = new ArrayList<String[]>();
24
```

Step 3. Test and build your codes

- Run “mvn test”

```
WoongheeLeeMacBookPro:~/git/TinySE-submit mvn test
```

```
[INFO] -------------------------------------------------------
[INFO]  T E S T S
[INFO] -------------------------------------------------------
[INFO] Running edu.hanyang.TokenizerTest
[WARNING] Tests run: 1, Failures: 0, Errors: 0, Skipped: 1, Time elapsed: 0.005 s - in edu.hanyang.TokenizerTest
[INFO] Running edu.hanyang.BPlusTreeTest
[WARNING] Tests run: 1, Failures: 0, Errors: 0, Skipped: 1, Time elapsed: 0.001 s - in edu.hanyang.BPlusTreeTest
[INFO] Running edu.hanyang.ExternalSortTest
time duration: 2627 msecs with 160 blocks of size 1024 bytes
[INFO] Tests run: 1, Failures: 0, Errors: 0, Skipped: 0, Time elapsed: 2.789 s - in edu.hanyang.ExternalSortTest
[INFO] Running edu.hanyang.QueryProcessTest
[WARNING] Tests run: 1, Failures: 0, Errors: 0, Skipped: 1, Time elapsed: 0 s - in edu.hanyang.QueryProcessTest
[INFO]
```

- Run “mvn package”

```
WoongheeLeeMacBookPro:~/git/TinySE-submit mvn package
```

```
WoongheeLeeMacBookPro:~/git/TinySE-submit ll target/
total 32
drwxr-xr-x  10 woonghee  staff   320B  3 29 14:28 .
drwxr-xr-x  13 woonghee  staff   416B  3 29 14:18 ..
-rw-r--r--   1 woonghee  staff    13K  3 29 14:18 2019123456-0.0.1-SNAPSHOT.jar
drwxr-xr-x   3 woonghee  staff    96B  3 29 13:15 classes
drwxr-xr-x   3 woonghee  staff    96B  3 18 10:31 generated-sources
drwxr-xr-x   3 woonghee  staff    96B  3 18 10:31 generated-test-sources
drwxr-xr-x   3 woonghee  staff    96B  3 29 14:15 maven-archiver
drwxr-xr-x   3 woonghee  staff    96B  3 18 10:31 maven-status
drwxr-xr-x  10 woonghee  staff   320B  3 18 10:31 surefire-reports
drwxr-xr-x   5 woonghee  staff   160B  3 29 13:35 test-classes
```

Updated

Step 4. Submit your module