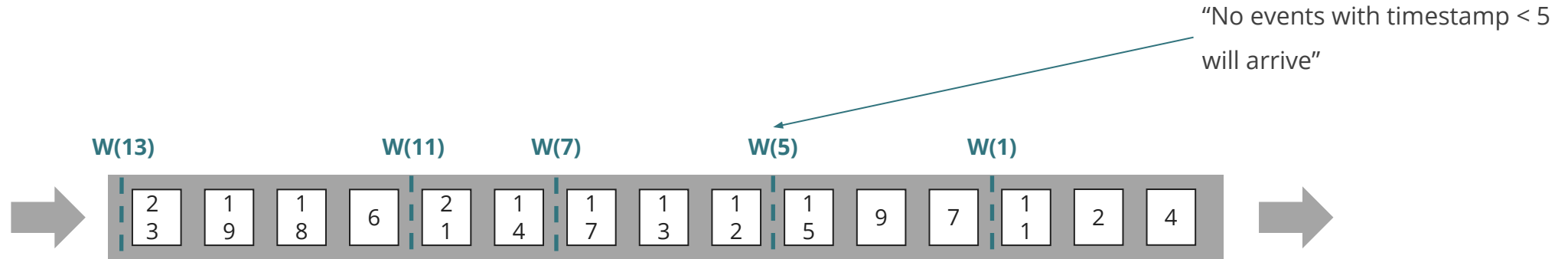


Watermarking

Konstantin Knauf, Solutions Architect

Watermarking

Reminder



- Watermarks push event time forward.
- They are provided by the data source or application
- They flow with the data stream and carry a timestamp

Watermarking

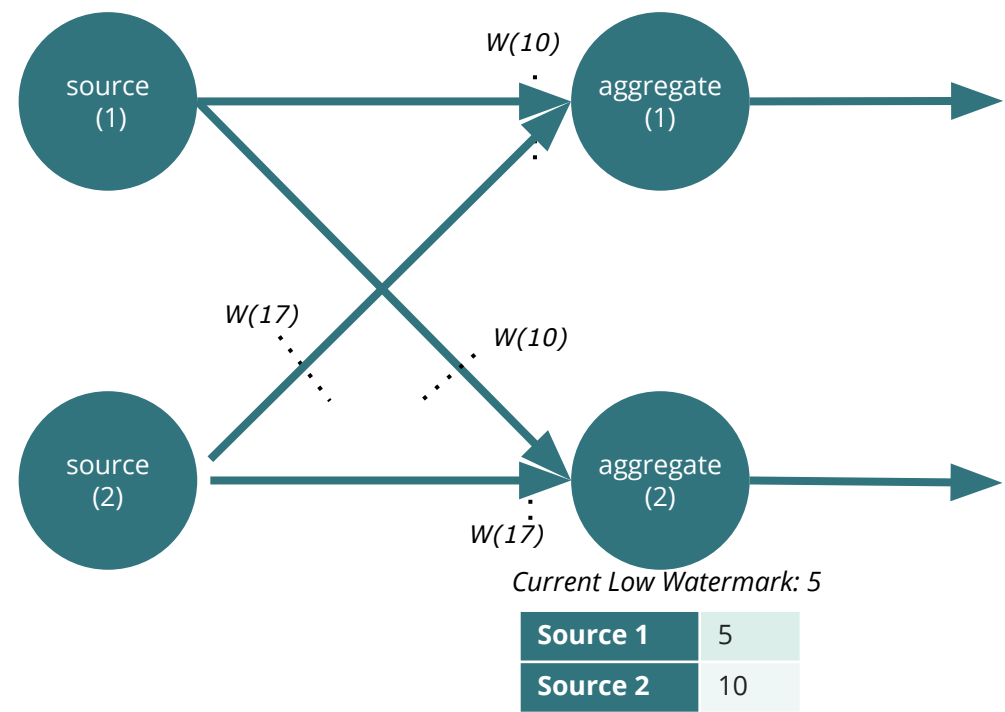
Watermark Assigners

- Periodic Watermarks
 - Based on a timer
 - `BoundedOutOfOrdernessGenerator` is an example
 - `ExecutionConfig.setAutoWatermarkInterval(msec)` controls the interval at which your periodic watermark generator is called
- Punctuated Watermarks
 - Based on something in the event stream



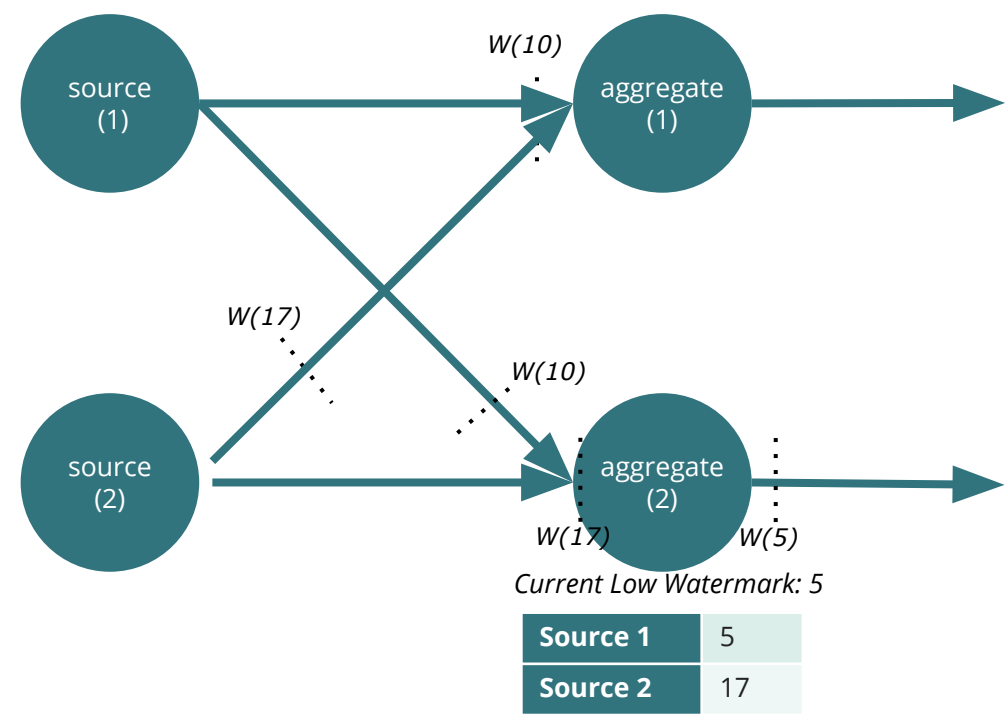
Watermarking

Watermarking in Parallel



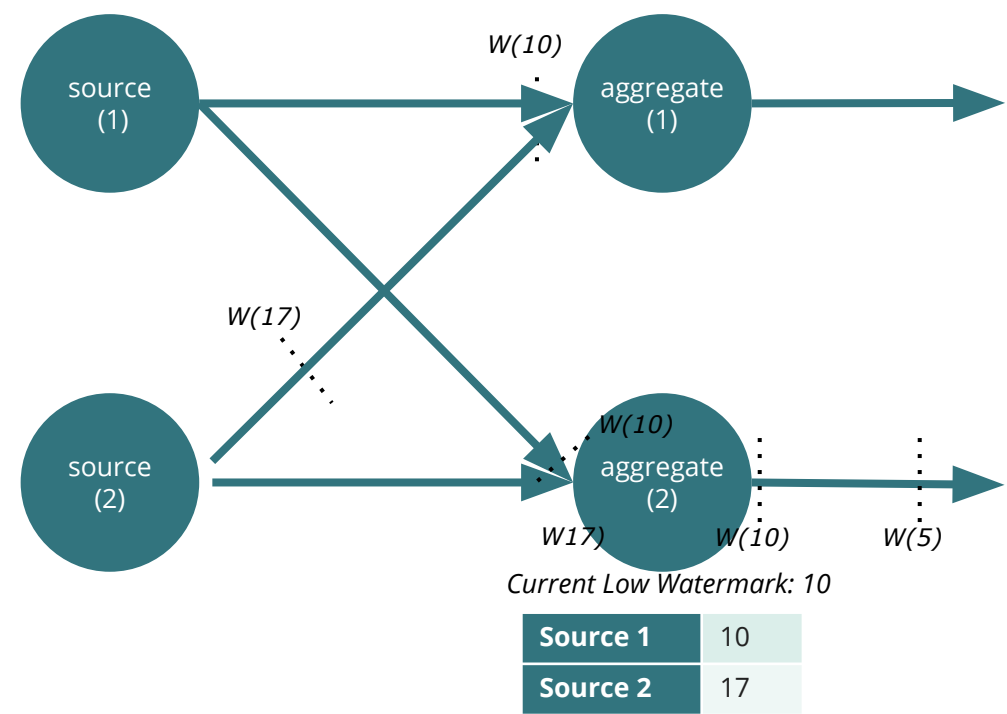
Watermarking

Watermarking in Parallel



Watermarking

Watermarking in Parallel



Advanced Topics



Kafka Per-Partition Watermarking

```
DataStream<KafkaEvent> input = env.addSource(new FlinkKafkaConsumer010<>(...))  
    .assignTimestampsAndWatermarks(new CustomWatermarkExtractor())
```

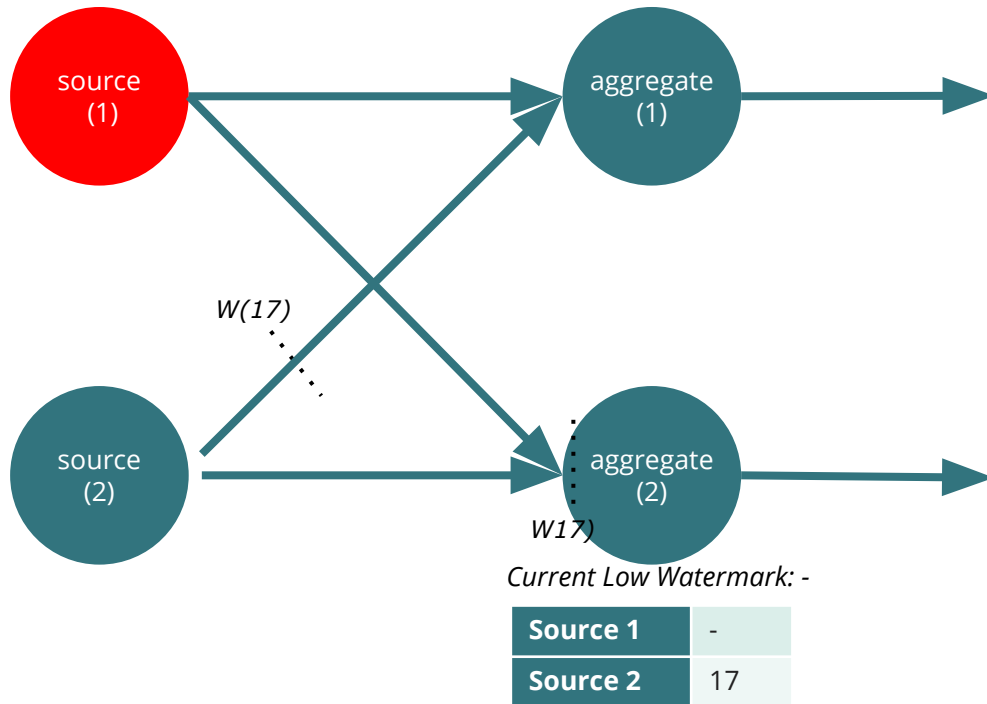
VS

```
DataStream<KafkaEvent> input = env.addSource(  
    new FlinkKafkaConsumer010<>(...).assignTimestampsAndWatermarks(new CustomWatermarkExtractor()))
```

- Watermarks are assigned per partition and aligned inside the source
- Per-Partition watermarks lead to better watermarking in certain situation



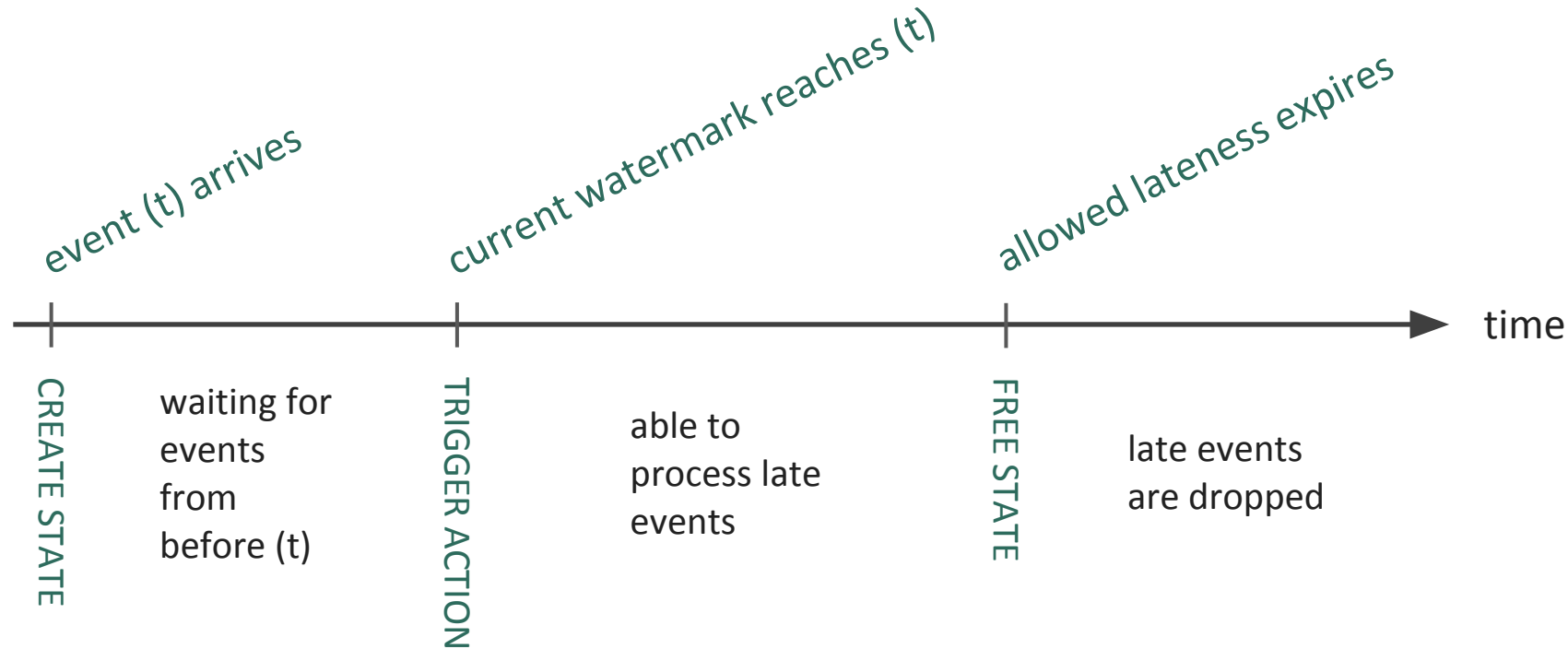
Idle Sources



Options

- Source marks itself idle via `SourceContext#markTemporarilyIdle()`
- `DataStream#rebalance()` prior to `TimestampAndWatermarkAssigner`
- `TimestampAssigner` or `SourceFunction` advance watermark based on processing time heuristic
 - Currently needs to be built by the user
 - FLINK-5479, FLINK-5017 will add support for Idle Timeouts in Flink

Dealing with Late Events



- Side outputs can be used to get a stream of the late data and deal with it
- Dealing with late data is application specific but the special code path could, e.g. update a value in the output database



ververica

konstantin@ververica.com

www.ververica.com

[@VervericaData](https://twitter.com/VervericaData)