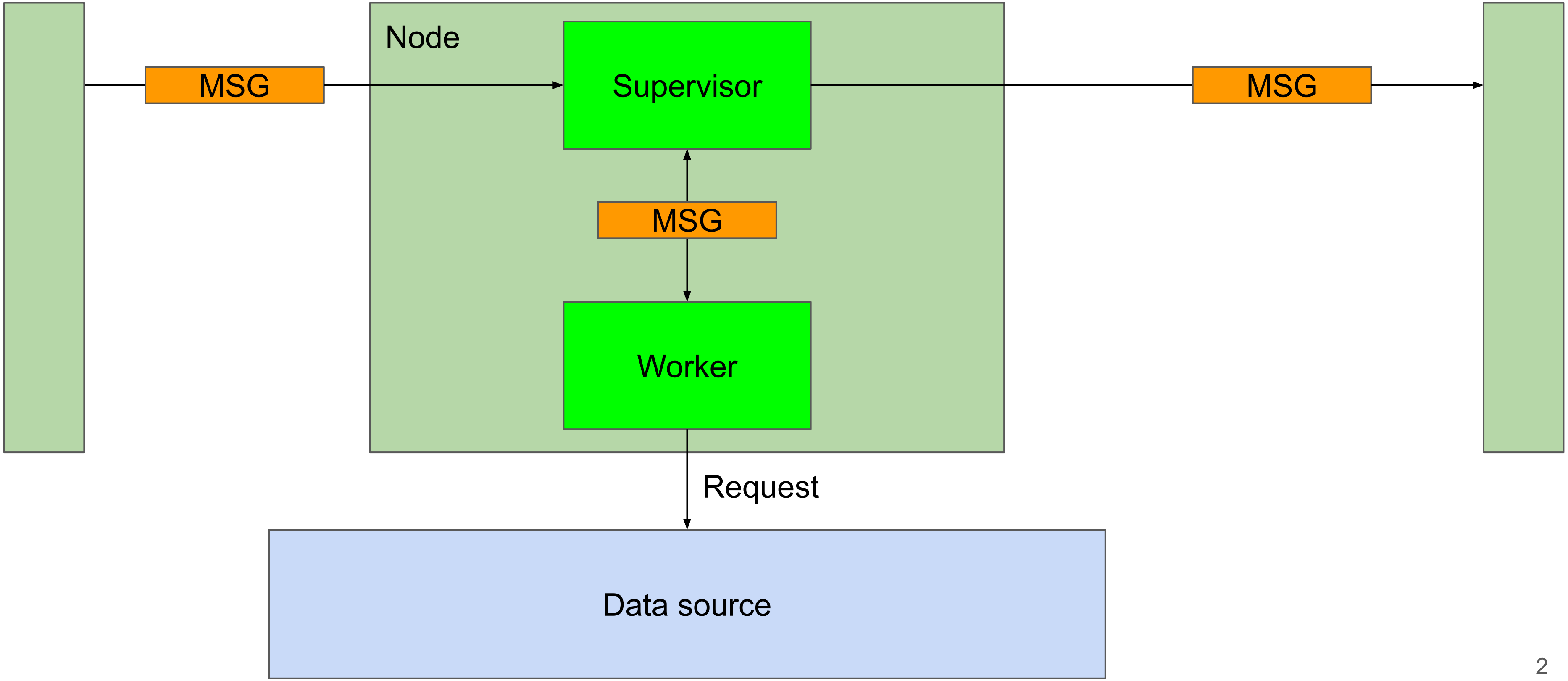
DKB Batch Processing

Vasilii Aulov

01.07.2021

Single message processing



Single message processing vs batch processing

Single: 1 message processed at a time, 1 request to each source per message.

- + Simpler.
- + Messages’ individuality makes it easier to deal with some errors.

Batch: batch is a temporary message aggregation that only exists in a given worker’s memory. 1 batch processed at a time, 1 request per batch.

- + Lower load on resources, both of DKB and sources.
- + Usage of sources’ existing methods of processing multiple records at once. AMI team specifically asked to improve this aspect of data4es.

Construction of batches

Supervisor-driven. Supervisor sends a number of messages to worker and then commands “start processing”.

+ Simpler (in terms of communication).

Worker-driven. Supervisor sends one message. Worker requests additional messages to construct a batch, processes it when deems it to be complete.

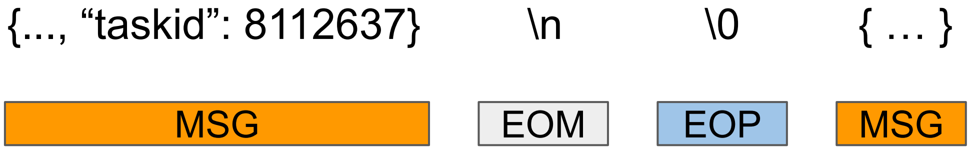
+ Worker (or, to be more precise, one who codes it) knows better: batch of what size it can optimally process at once.

− Large batch size can lead to excessive consumption of resources.

Communication between worker and supervisor

- **<u>Signal-based</u>**

Further development on top of existing implementation. Communication by means of signals (markers): special symbols or symbol combinations.



{..., “taskid”: 8112637} \n \0 { … }

MSG EOM EOP MSG

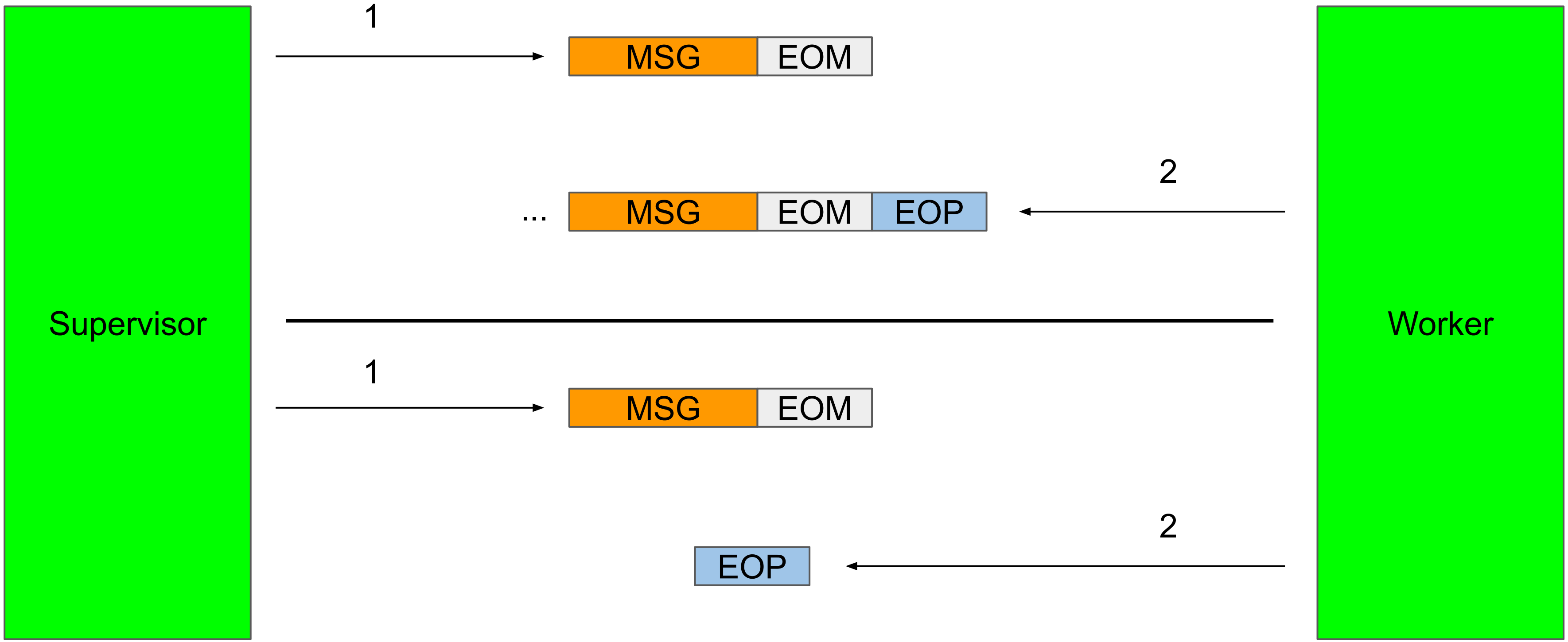
Markers:
EOM - End Of Message
EOP - End Of Process

BNC - Batch Not Complete
EOB - End Of Batch

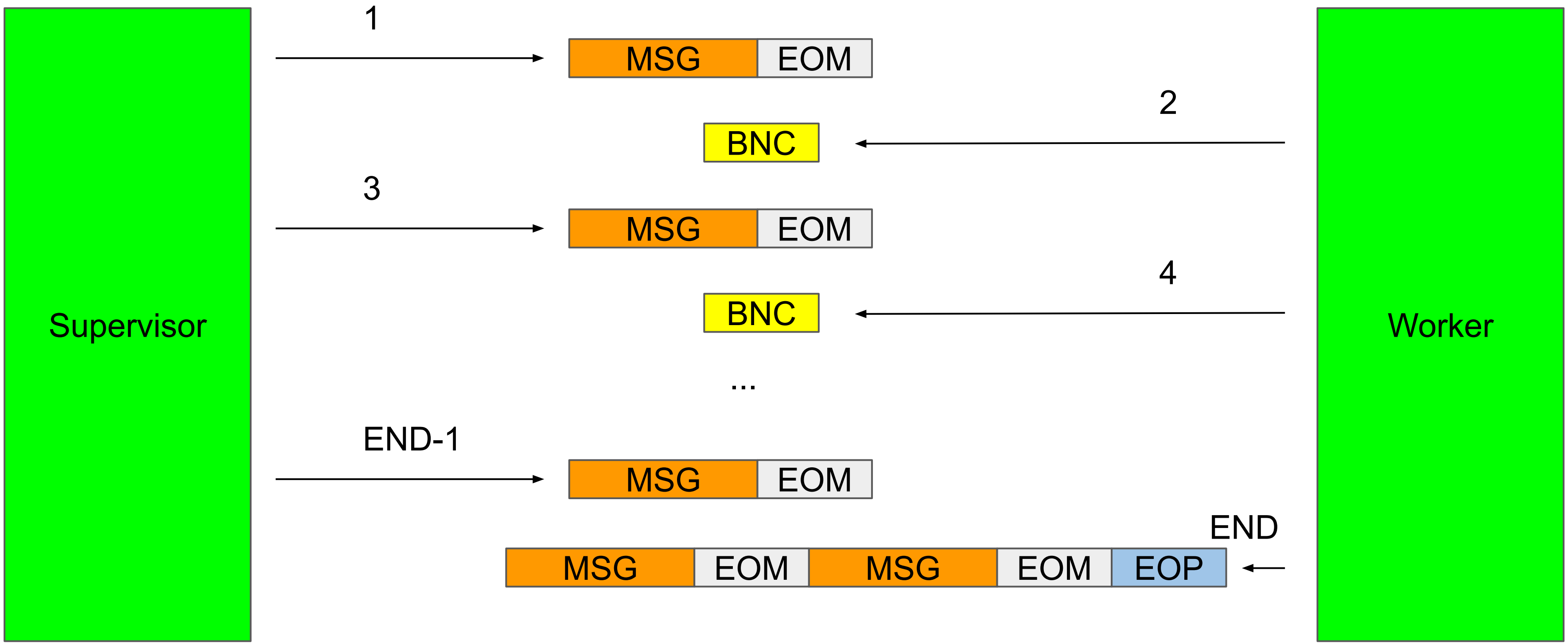
- Protocol-based

Rework of existing implementation to use messages with headers (HTTP, STOMP).

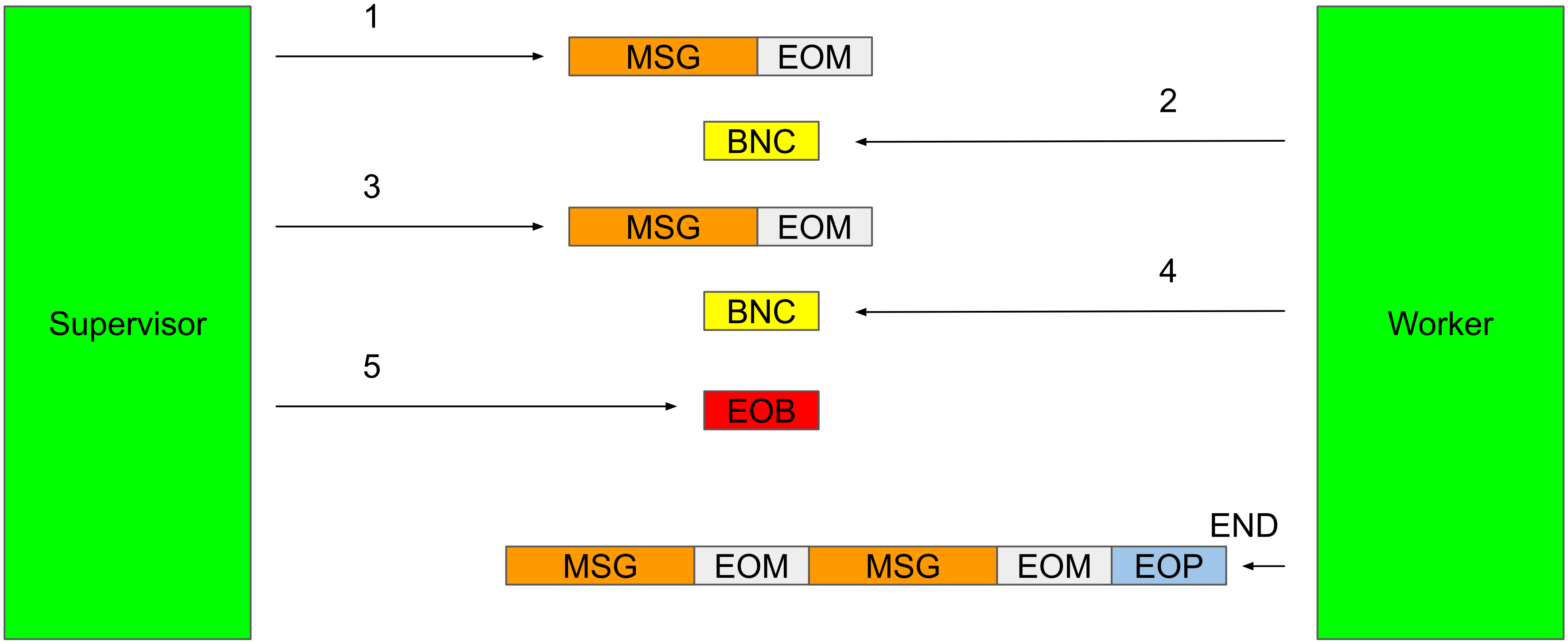
Communication: single message mode



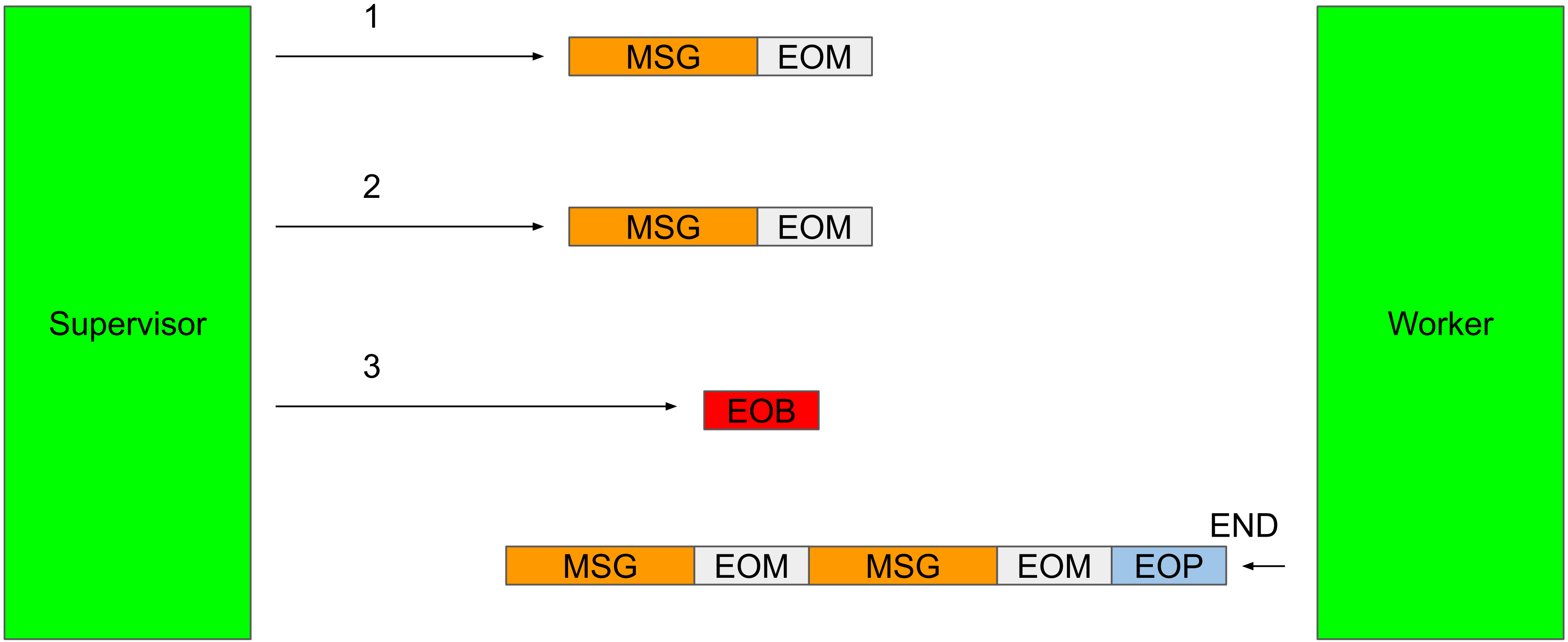
Communication: worker-driven batch mode



Communication: EOB



Communication: supervisor-driven batch mode



Technical details

- custom_readline(): marker and message, separate or together?
- InputStream: how to handle unknown markers?
- ProcessorStage: handling BNC when Producer switches output.
- What happens in supervisor-driven mode when the stage does not support batch processing?