Flink

Table | Gelly

CEP | Flink ML

DataStream API | DataSet API

Runtime

Local
Single JVM

Cluster
Standalone, YARN

Cloud
GCE, EC2

Libraries

High level APIs

Flink

Table

Gelly

Libraries

CEP

Flink ML

DataStream API

DataSet API

High level APIs

Runtime

Local

Single JVM

Cluster

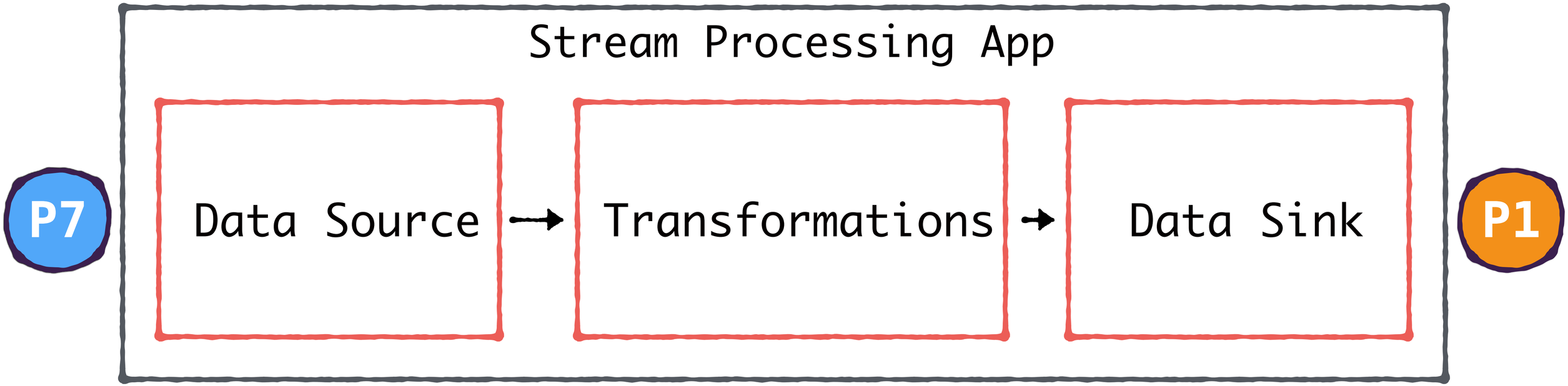
Standalone, YARN

Cloud

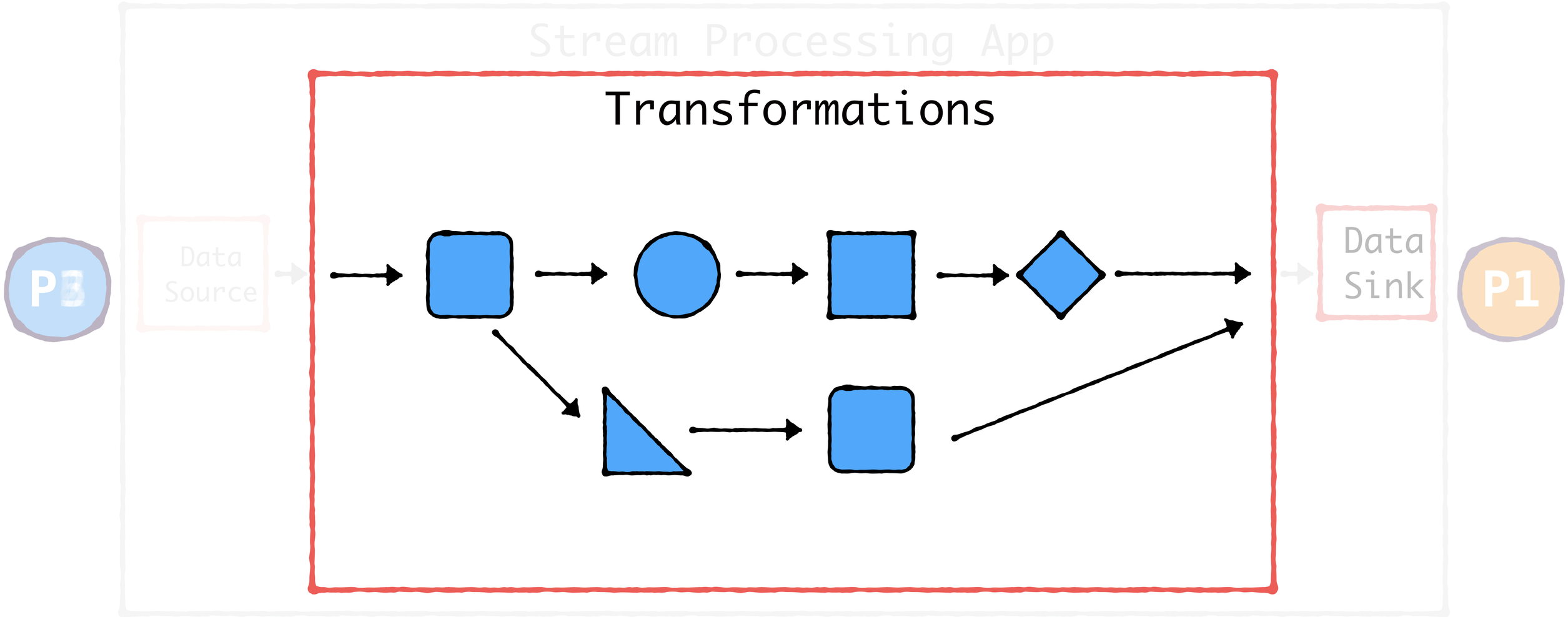
GCE, EC2

Flink Programming Model

Flink Programming Model



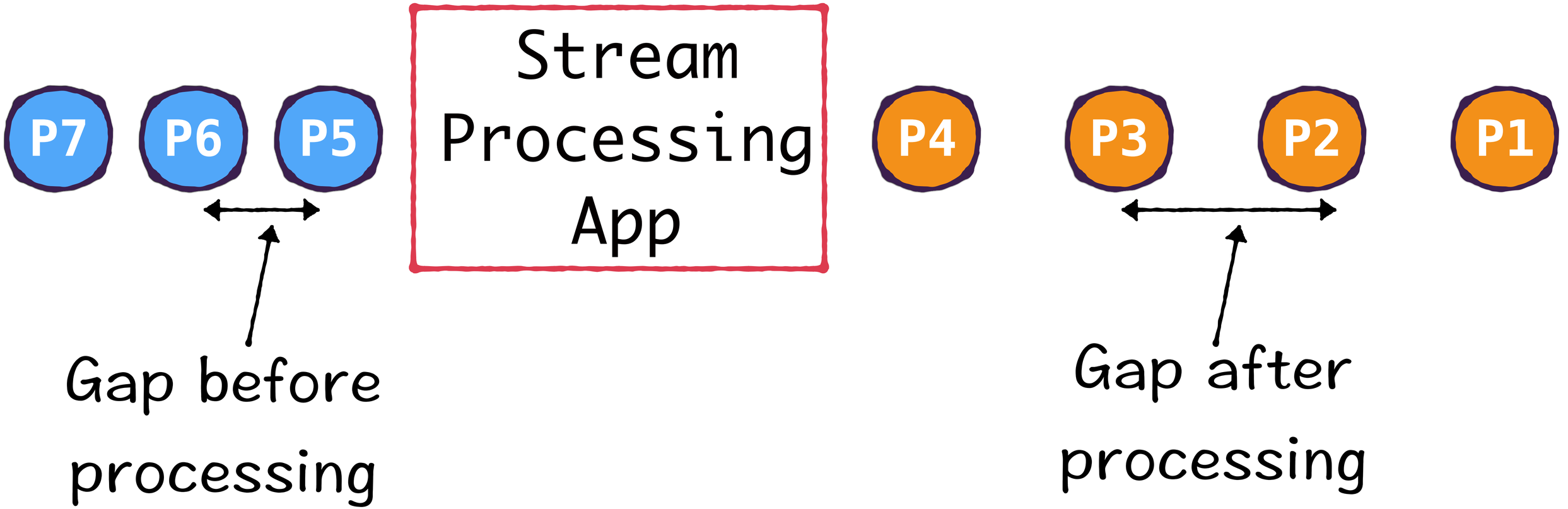
Flink Programming Model



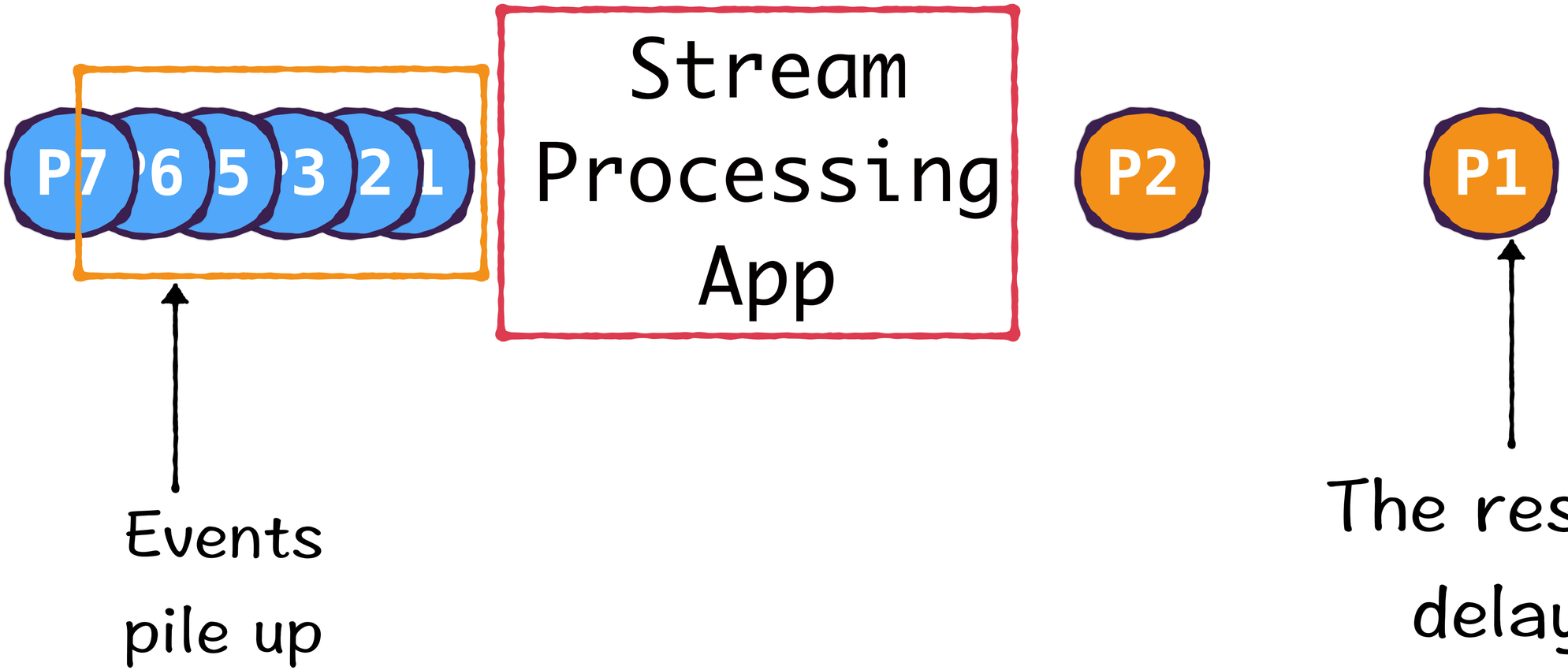
Stream processing

Process events fast with no loss of information

Gap increased during processing time

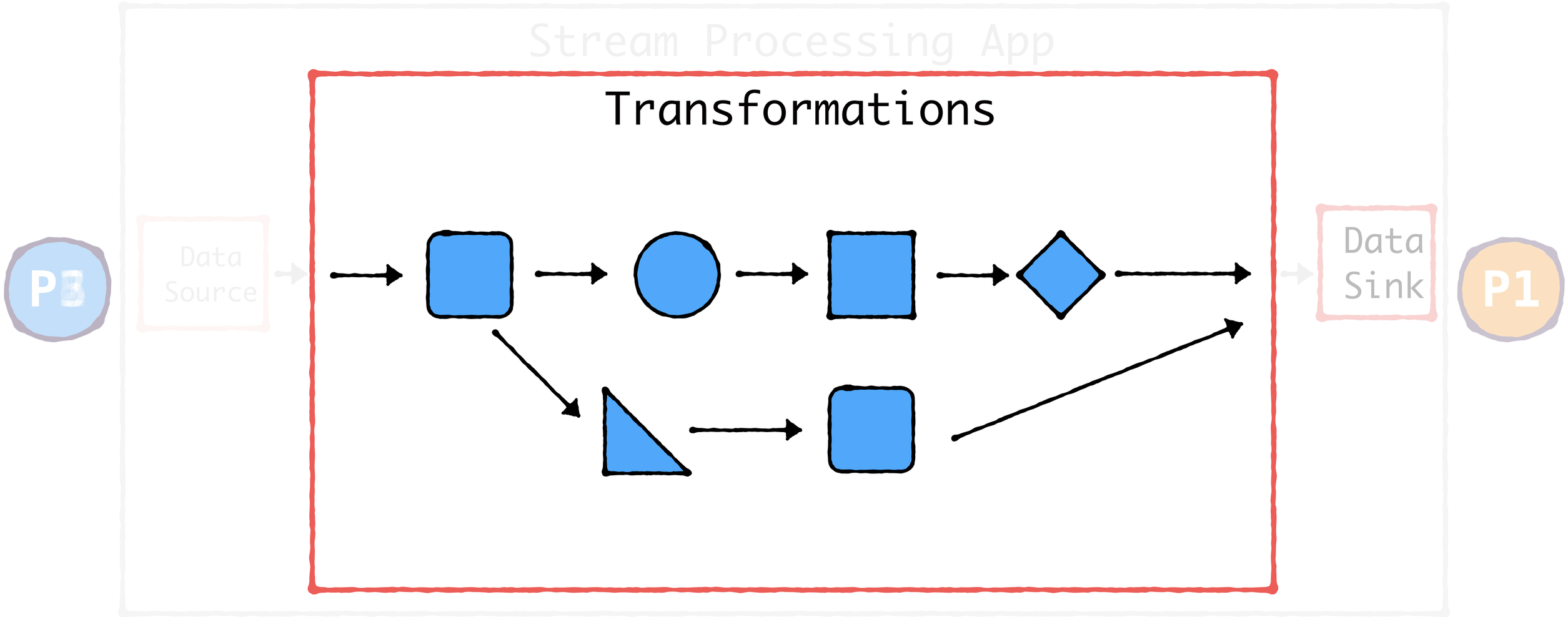


If processing time >> event arrival rate

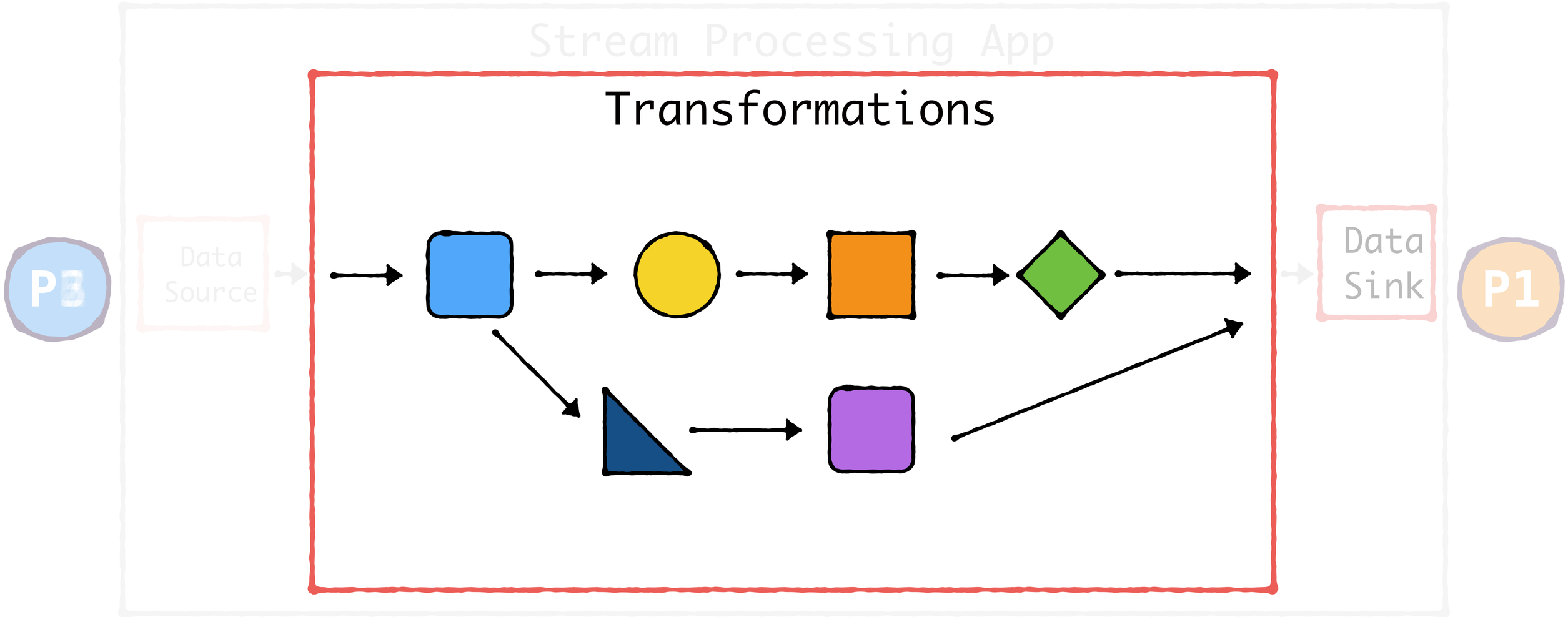


How do we reduce processing time?

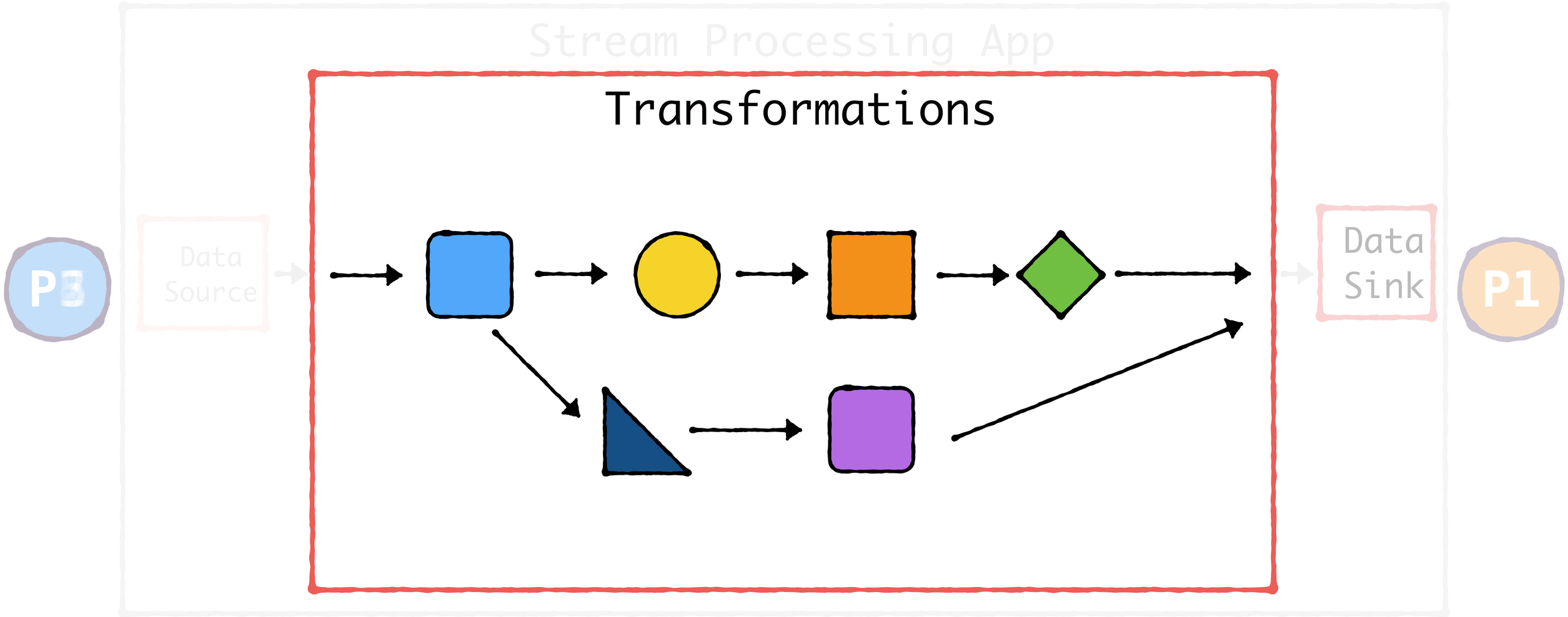
Each operator performs a small transformation



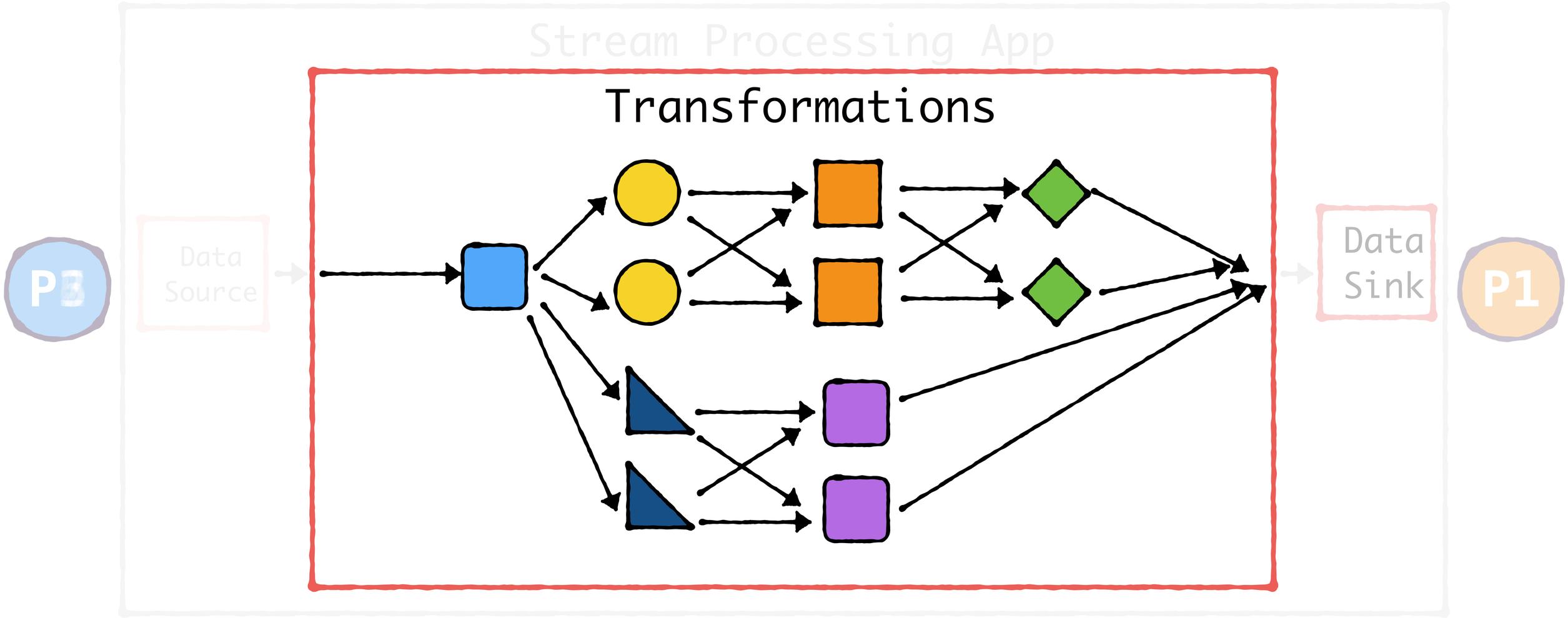
Each operator performs a small transformation



Add multiple instances of each operator

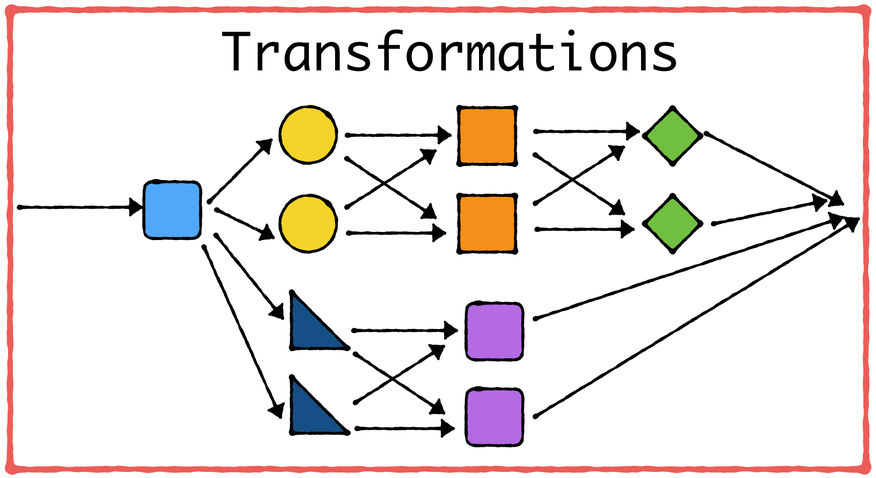


Add multiple instances of each operator



Parallelism in Flink

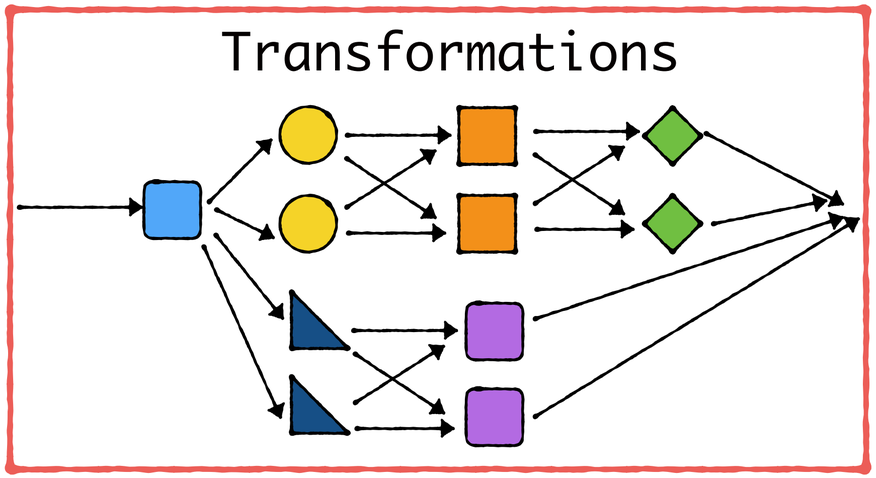
A directed acyclic graph



Users express transformations

Flink translates to a DAG of operators

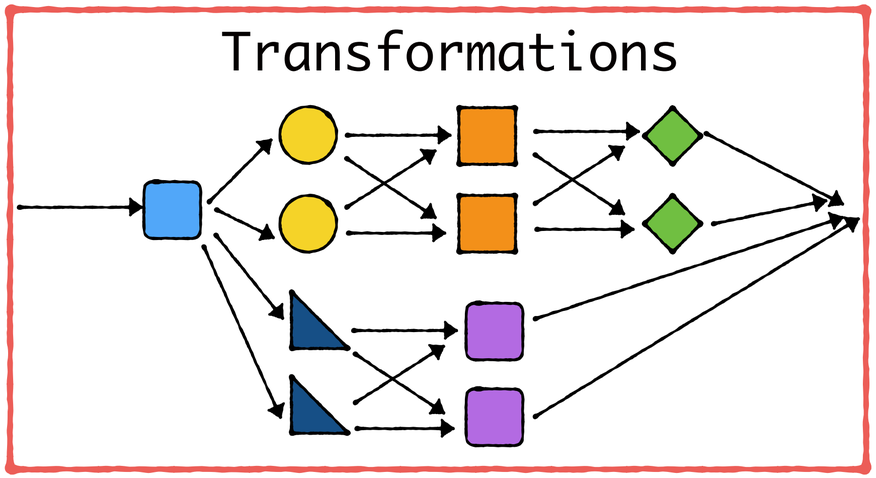
Parallelism in Flink



Flink translates to
a DAG of operators

Groups operators
together into tasks

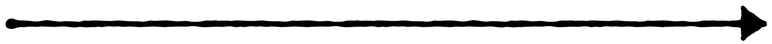
Parallelism in Flink

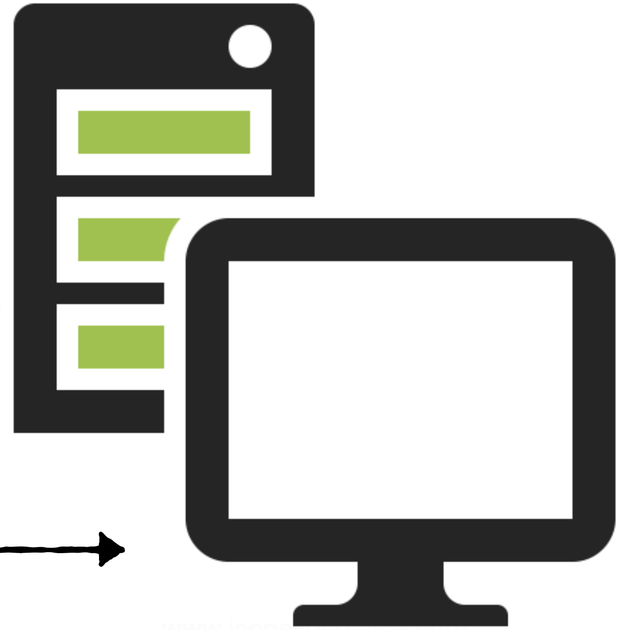


Groups operators together into tasks

Tasks run in parallel on separate threads

Flink architecture

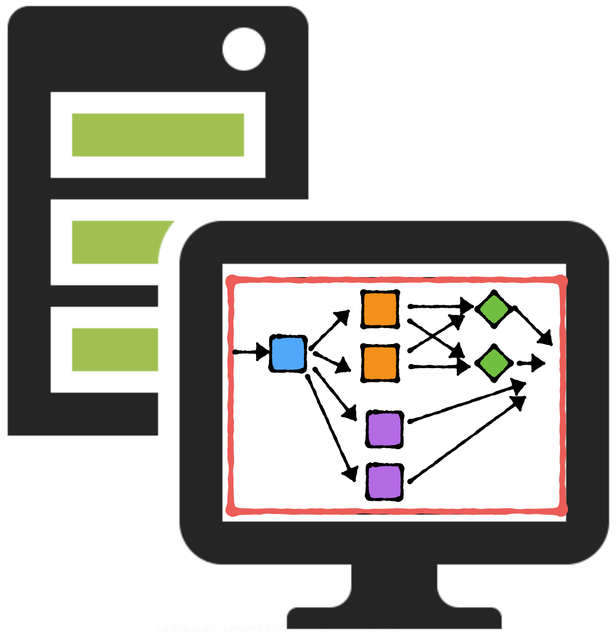




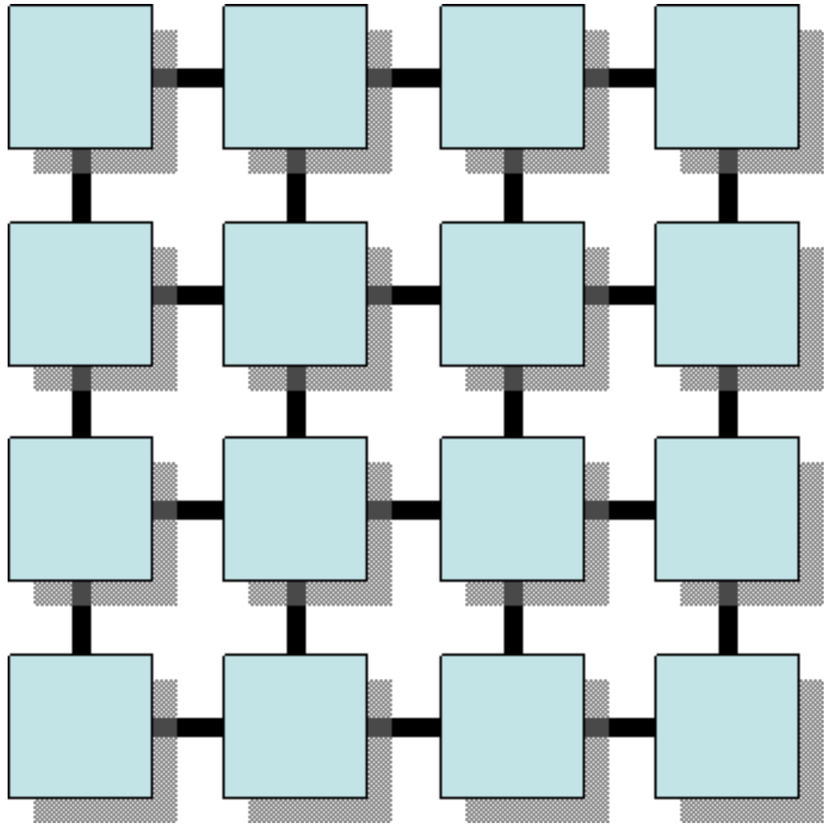
User

Flink client

Flink architecture



Flink Client



Flink Cluster

Flink architecture

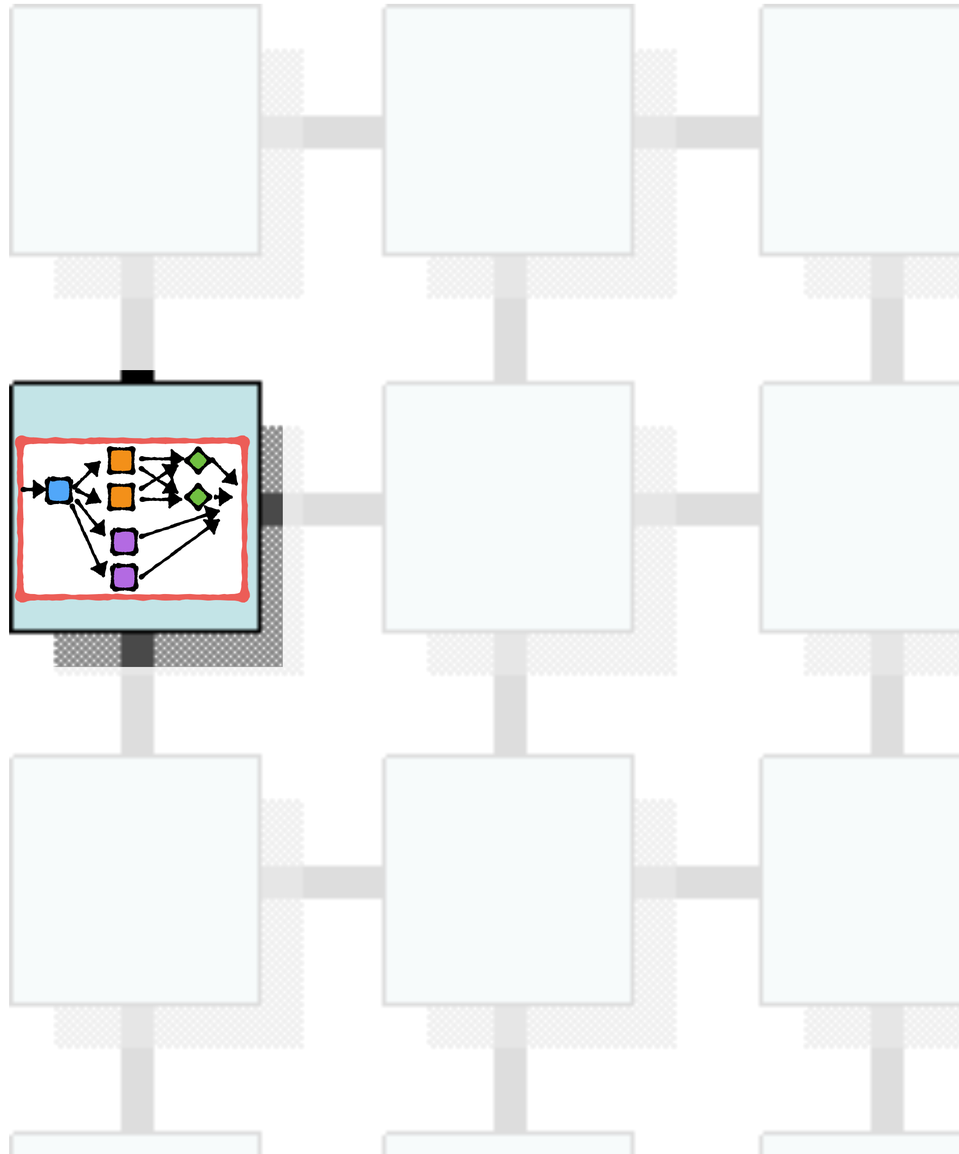
Flink Master

Central co-ordinating process

Schedules tasks

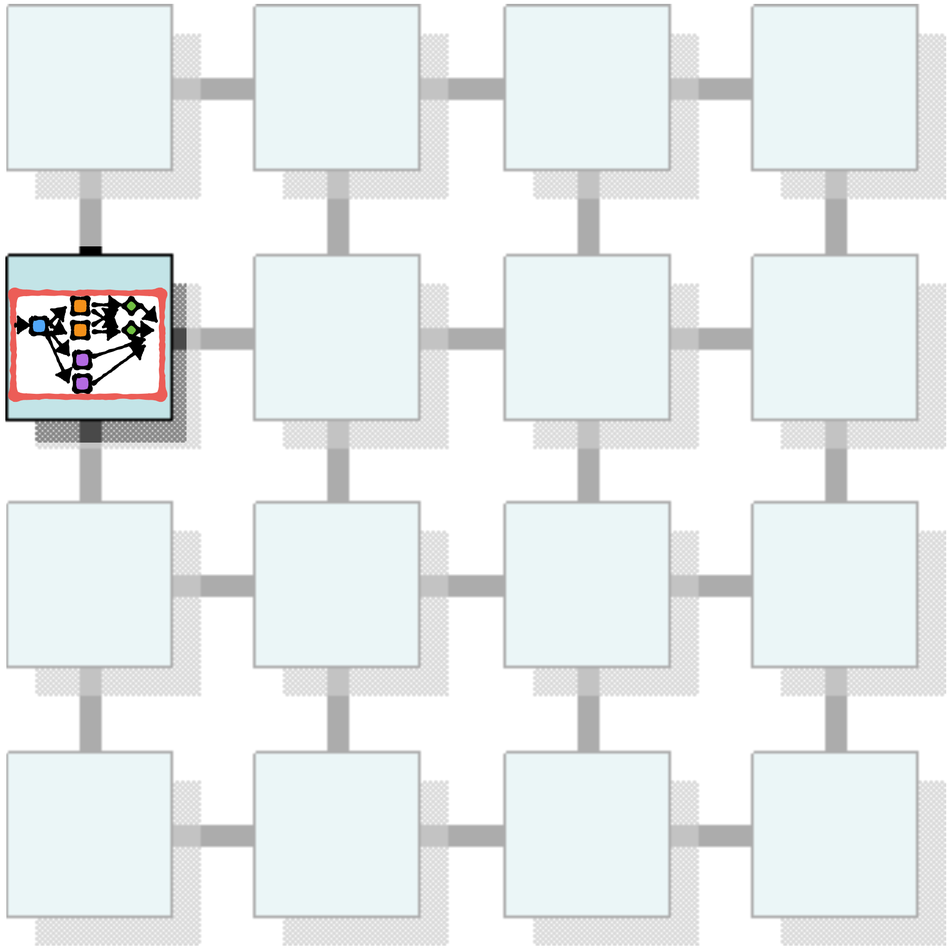
JobManager

(Leader and standby)



Flink architecture

Tasks get assigned
to workers

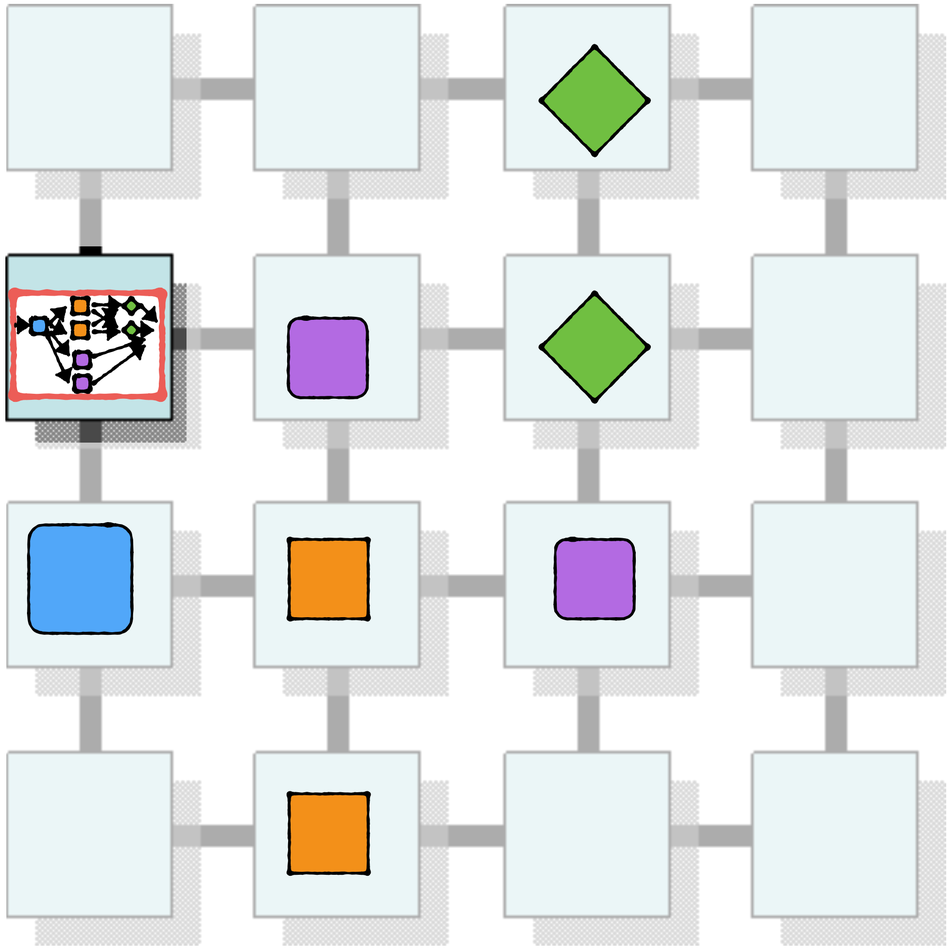


Flink architecture

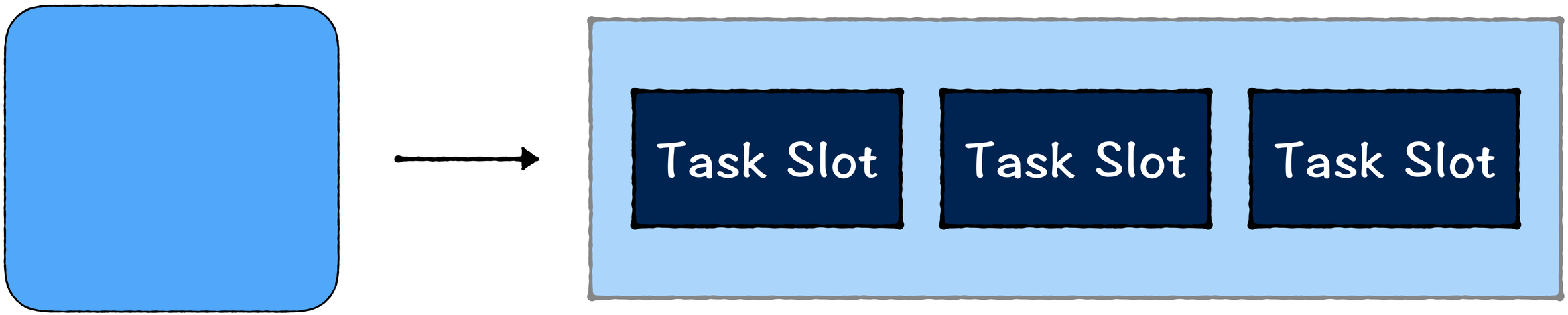
Tasks get assigned
to workers

TaskManagers

Run sub-tasks on
separate threads



Flink architecture



Do not compete with
each other for resources

Can share data

Before Starting

Java, Java frameworks

Jars, compiling code and debugging in Java

Maven to build projects