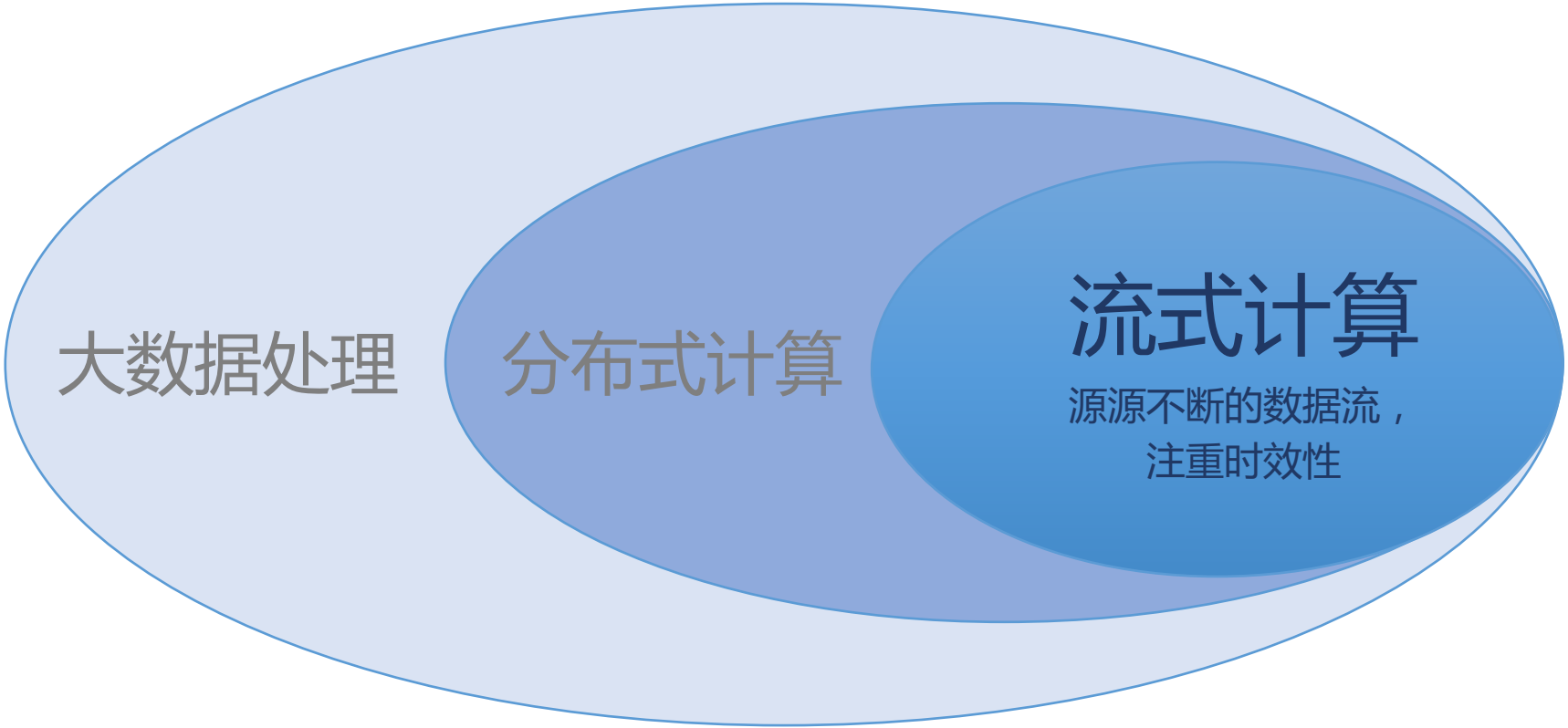
百度DStream3

百度 程怡
2018.10

流式计算是什么？

流式计算系统



流式计算系统

DStream1.0
纯流式计算系统
2011

TM1.0
小批量流式计算系统
2011

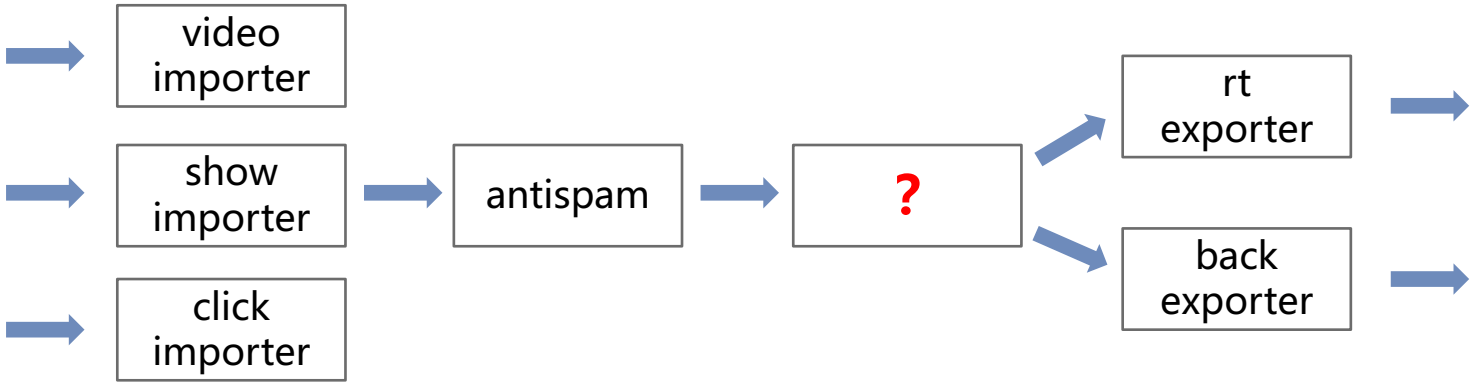
DStream3.0
新一代纯流式计算系统
2017

Hadoop
批量计算系统
2008

Bigpipe
消息传输系统
2010前后

百度DStream3 流式计算系统，
致力于解决 大规模、低延迟、高准确、高可用 的
无限数据流计算问题

Motivation



CLC(Correctness-Latency-Cost)

Motivation


Latency
搜索建库
反作弊
数仓ETL
用户实时模型
监控&报警
广告风控
账单报表
广告计费
Cost
Correctness

DStream3

- **End-to-end 3合1 Once语义**
- 3合1 Low Watermark
- Pull & Long Polling RPC
- multi-distribution mode
- 10秒级保活 & 防僵尸机制
- **泄洪机制**
- 面向容器，多租户隔离

- **Upstream-dependent mode**
- 用户多线程支持
- micro-bundle支持
- **动态 Upgrade/DAG/parallel/resource**
- **自动反压 & 主动限速**
- 框架避让用户线程
- 资源、权限管理

- 多层次SDK
- 多语言支持
- metric
- profiling
- tracing
- 报警
- log

Once语义

- At-most-once：最多只处理一次，可能丢数据
- At-least-once：至少处理一次，可能重复处理
- Exactly-once：处理且仅处理一次，不重不丢
- End-to-end Exactly-once：与上下游协同，保证最终结果准确性

Once语义

- At-most-once：最多只处理一次，可能丢数据
- At-least-once：至少处理一次，可能重复处理
- Exactly-once：处理且仅处理一次，不重不丢
- End-to-end Exactly-once：与上下游协同，保证最终结果准确性

Exactly-once

备份 + 重放

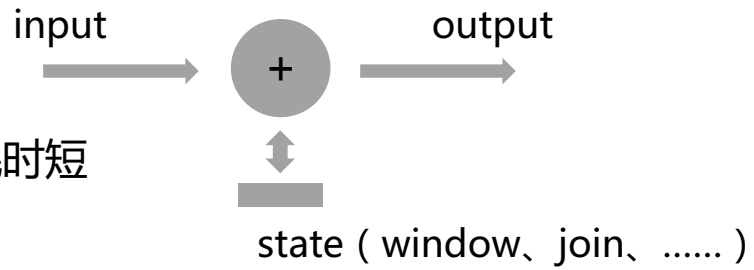
At-least-once

标记 + 去重

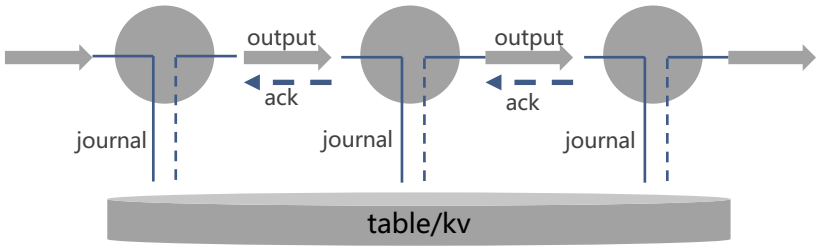
Exactly-once

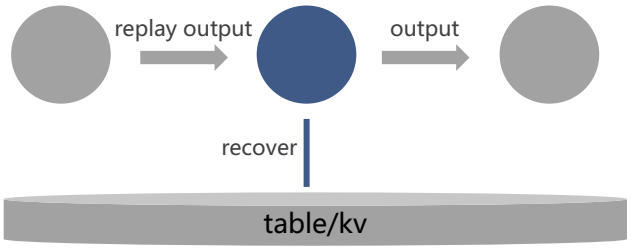
备份

- 乐观备份：
  - 假设执行环境很好，故障极少发生
  - 较低频率持久化state，不持久化output。一般是全量
  - runtime latency低，持久化成本低，failover耗时长
- 悲观备份：
  - 假设执行环境较为恶劣，故障频繁发生
  - 同时持久化state和output，一般是增量
  - runtime latency高，持久化成本高，failover耗时短
- DStream3：
  - 悲观备份：storage-dependent mode
  - 乐观备份：upstream-dependent mode



Storage-dependent mode





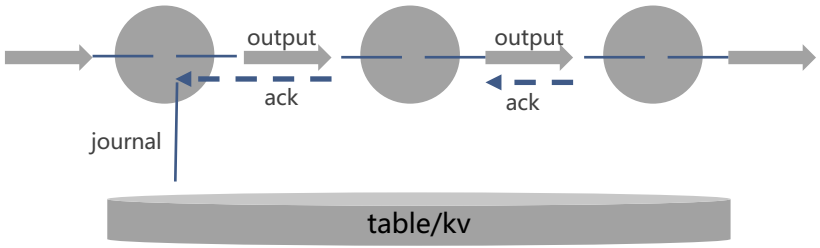
1. receive tuple from upstream, or timer expired
2. call user’ s on_tuple()/on_timer()
3. user emit output, and/or modify state, and/or modify timer
4. sync journal all changes of step3 atomically (can be bundled)
5. output can be pull by downstream

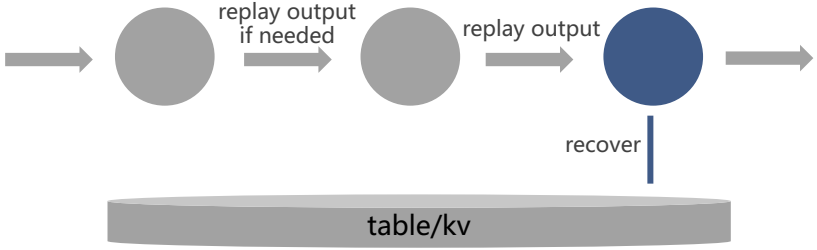
1. recover pending output, timer and state from journals
2. replay output normally
3. receive tuple from upstream, dedup if Exactly-once

Storage-dependent mode

CLC(Correctness-Latency-Cost) vs Storage-dependent mode

Upstream-dependent mode





- 受限场景开启
- 仅增量journal state
- 慢ack自动触发storage-dependent journal切换

- 外存恢复state
- 逐级溯源恢复output

output

state

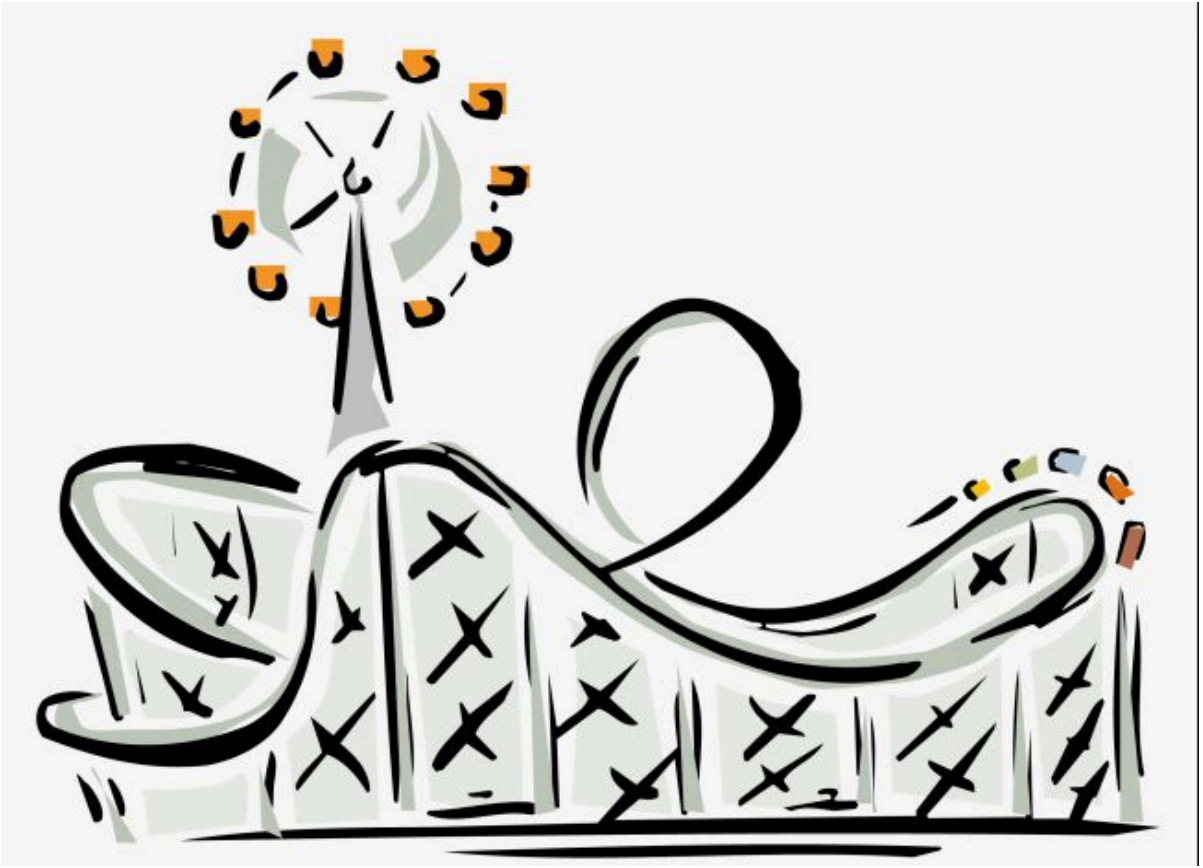
Exactly-once

Storage-dependent mode vs Upstream-dependent mode

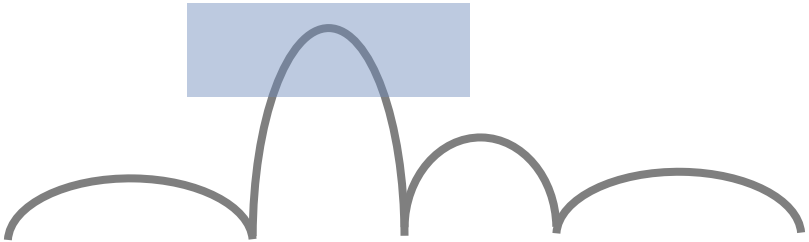
o(100ms) vs o(10ms)

无对齐点，全自治

Reality

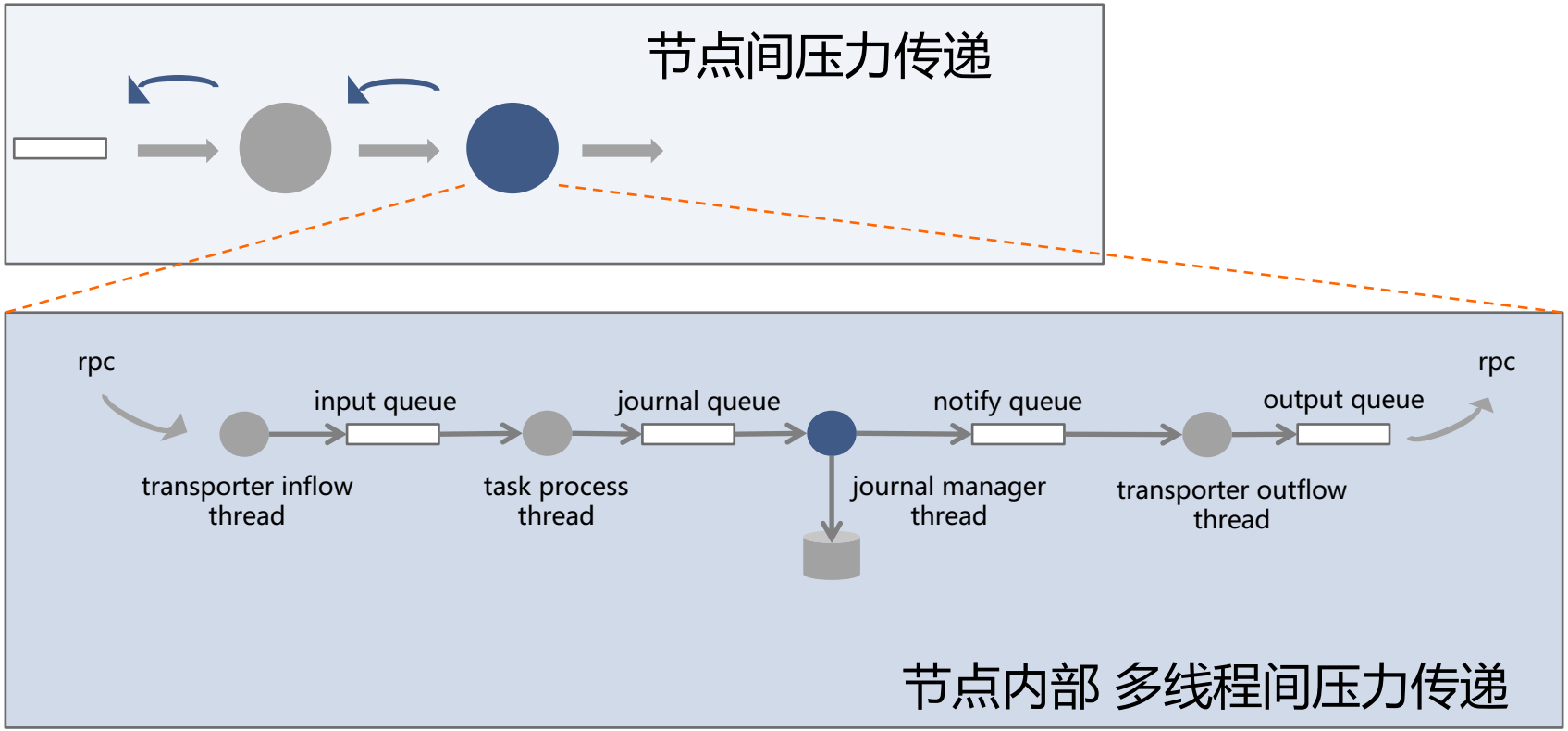


流量波动



- 排队
- 反压流控
- 拒绝流量
- 主动泄洪
- 服务降级
- 动态更新
- 扩容
- 动态扩容

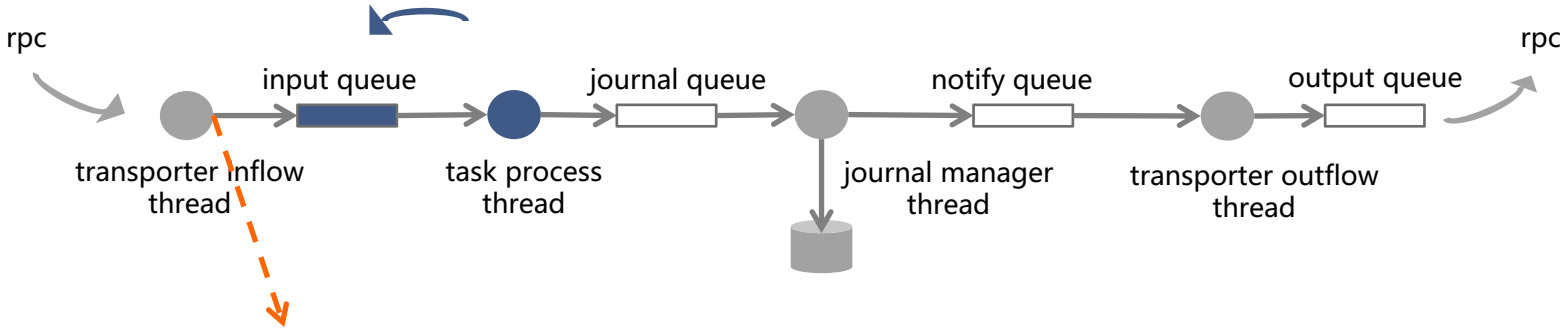
Back-pressure



Back-pressure

CLC(Correctness-Latency-Cost) vs Back-pressure

Drop

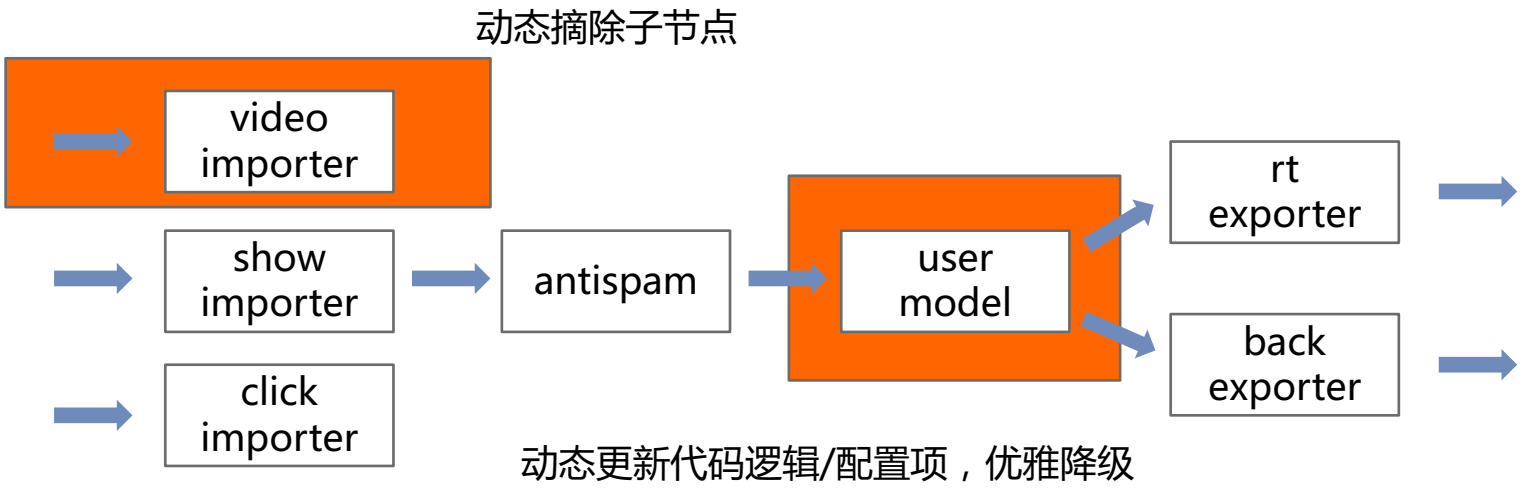


- queue full : drop
- high watermark : slow down request

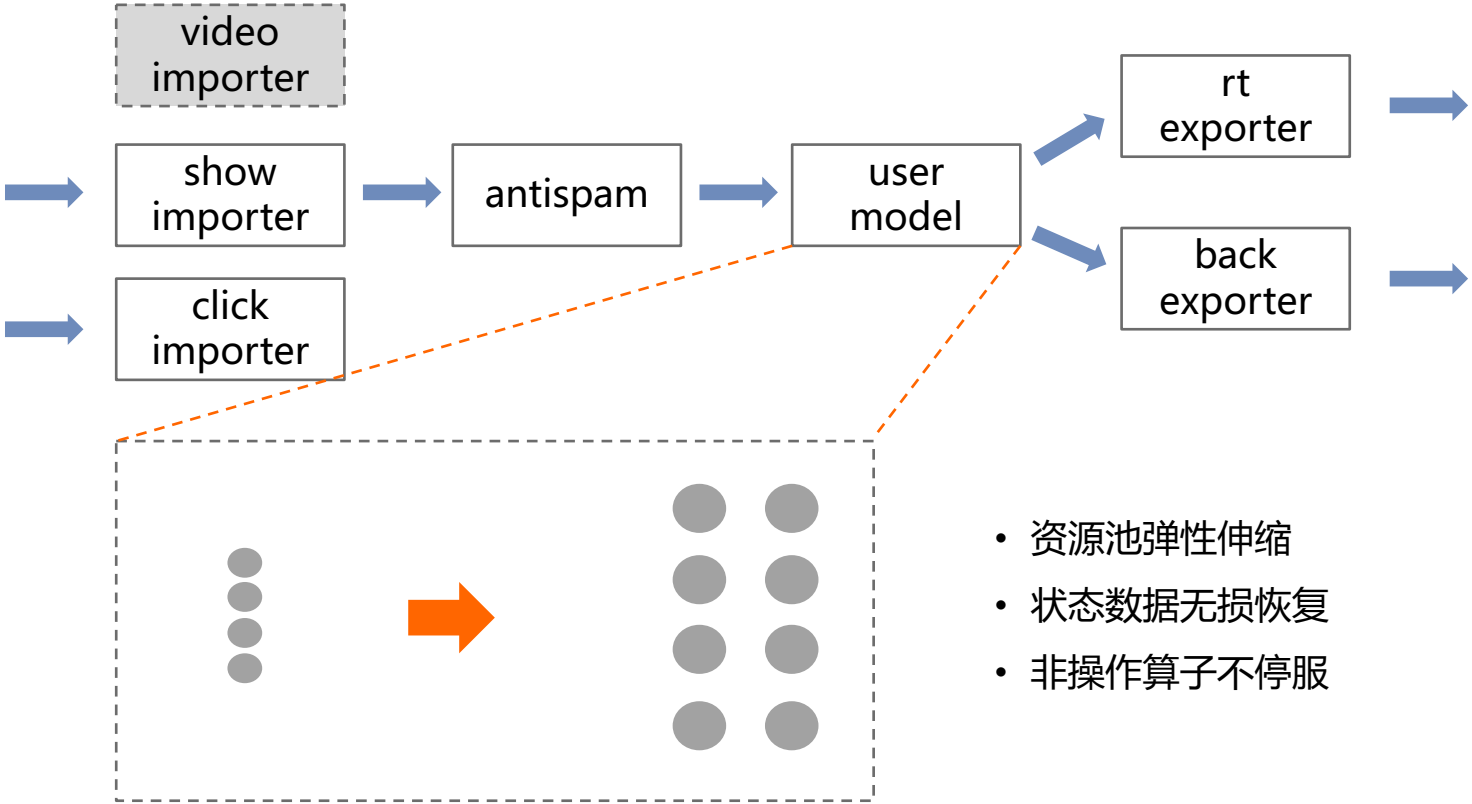
Drop

CLC(Correctness-Latency-Cost) vs Drop

Dynamical


动态摘除子节点
video importer
show importer
click importer
antispam
user model
rt exporter
back exporter
动态更新代码逻辑/配置项，优雅降级

Parallel & Resource



- 资源池弹性伸缩
- 状态数据无损恢复
- 非操作算子不停服

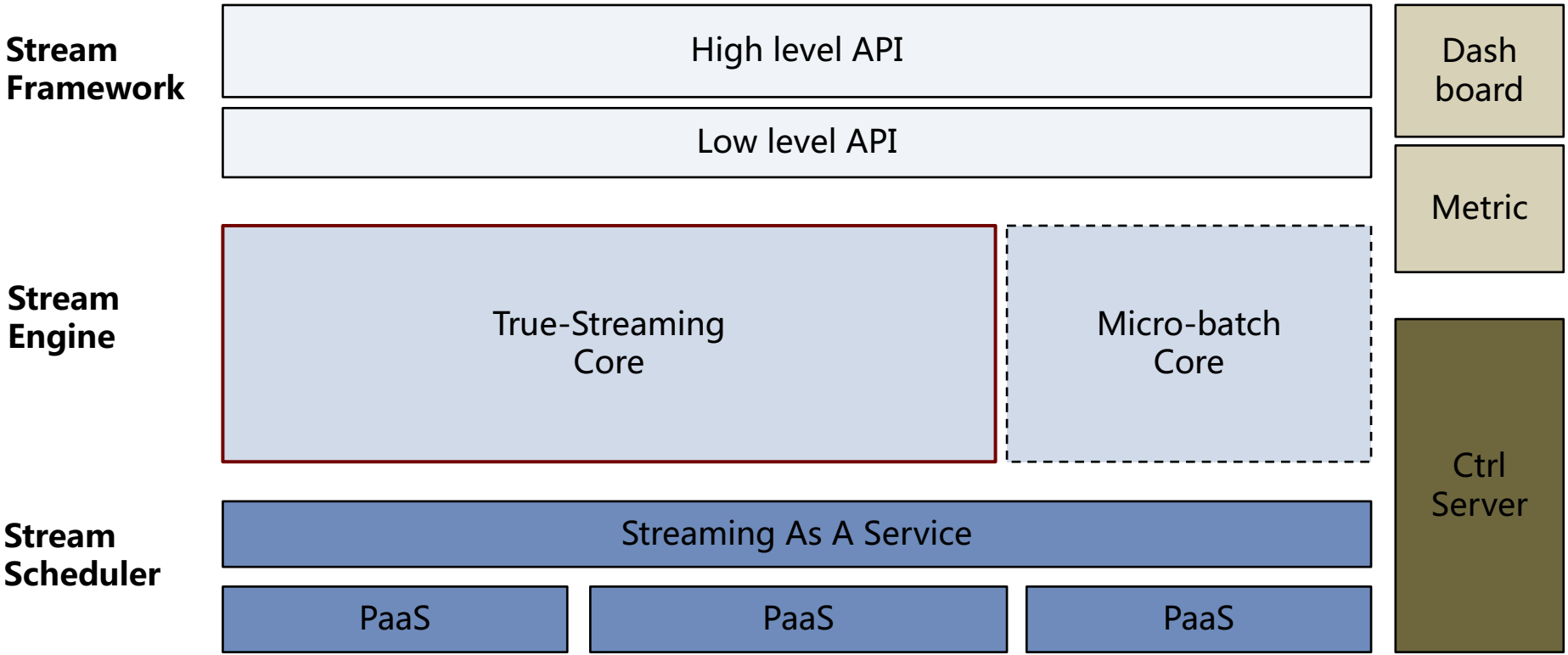
DStream3

- **End-to-end 3合1 Once语义**
- 3合1 Low Watermark
- Pull & Long Polling RPC
- multi-distribution mode
- 10秒级保活 & 防僵尸机制
- **泄洪机制**
- 面向容器，多租户隔离

- **Upstream-dependent mode**
- 用户多线程支持
- micro-bundle支持
- **动态 Upgrade/DAG/parallel/resource**
- **自动反压 & 主动限速**
- 框架避让用户线程
- 资源、权限管理

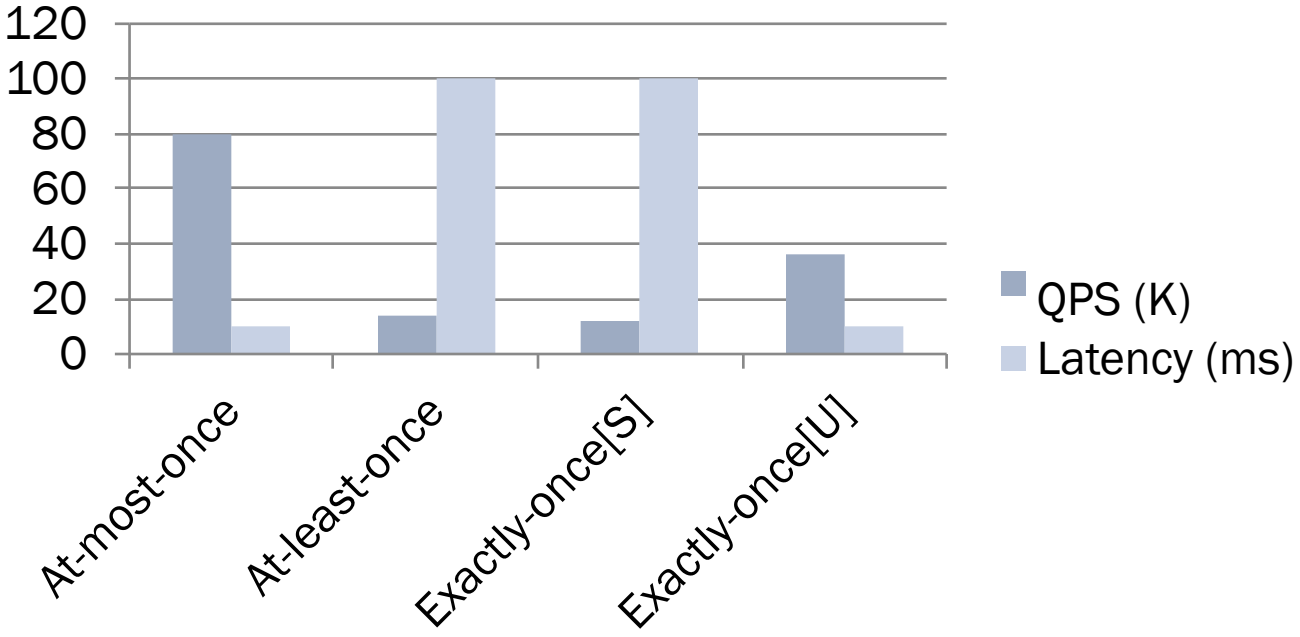
- 多层次SDK
- 多语言支持
- metric
- profiling
- tracing
- 报警
- log

Arch


Stream Framework
High level API
Low level API
Dash board
Metric
Stream Engine
True-Streaming Core
Micro-batch Core
Ctrl Server
Stream Scheduler
Streaming As A Service
PaaS
PaaS
PaaS

Perf

- tuple size = 1KB, 3归一核 container(约等于2core)



Next

- Robust
- Skew (data/computing/host/network)
- AI+
- Dual-core

ArchSummit 全球架构师峰会 [北京]

主办方 Geekbang 极客邦科技 InfoQ

技术创新的浪潮接踵而来，继续搬砖还是奋起直追？

云数据 | AI | 区块链 | 架构优化

高效运维 | CTO技术选型 | 微服务 | 新开源框架

会议：2018年12月07-08日 培训：2018年12月09-10日

地址：北京·国际会议中心

Q&A