RISELab: Enabling Intelligent Real-time Decisions

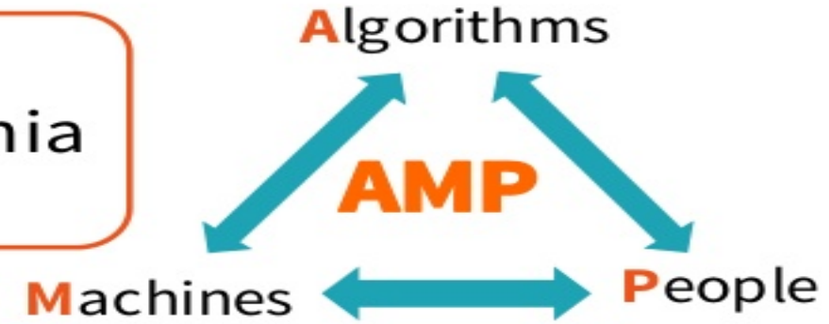
Ion Stoica

February 8, 2017

Berkeley's AMPLab (2011-2016)

Goal: Next generation of open source data analytics stack for industry & academia

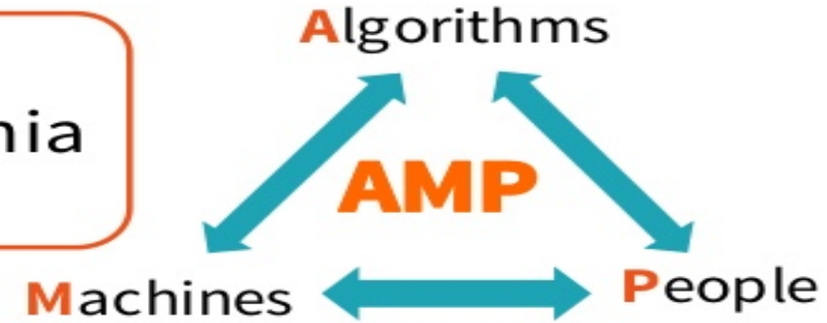
Berkeley Data Analytics Stack (BDAS)



Berkeley's AMPLab (2011-2016)

Goal: Next generation of open source data analytics stack for industry & academia

Berkeley Data Analytics Stack (BDAS)



Apache MESOS

ALLUXIO

• • •

RISE: Real-time Intelligent Secure Execution

RISELab

From live data to real-time decisions

AMPLab

From batch data to advanced analytics

RISE Lab (2017-2022)

12 faculty across AI, systems, security, and architectures

11 Founding sponsors

蚂蚁金服 ANT FINANCIAL · amazon web services · Capital One · ERICSSON · GE · Google · HUAWEI · intel · IBM · Microsoft · vmware

Why?

Data only as valuable as the decisions it enables

Why?

Data only as valuable as the decisions it enables

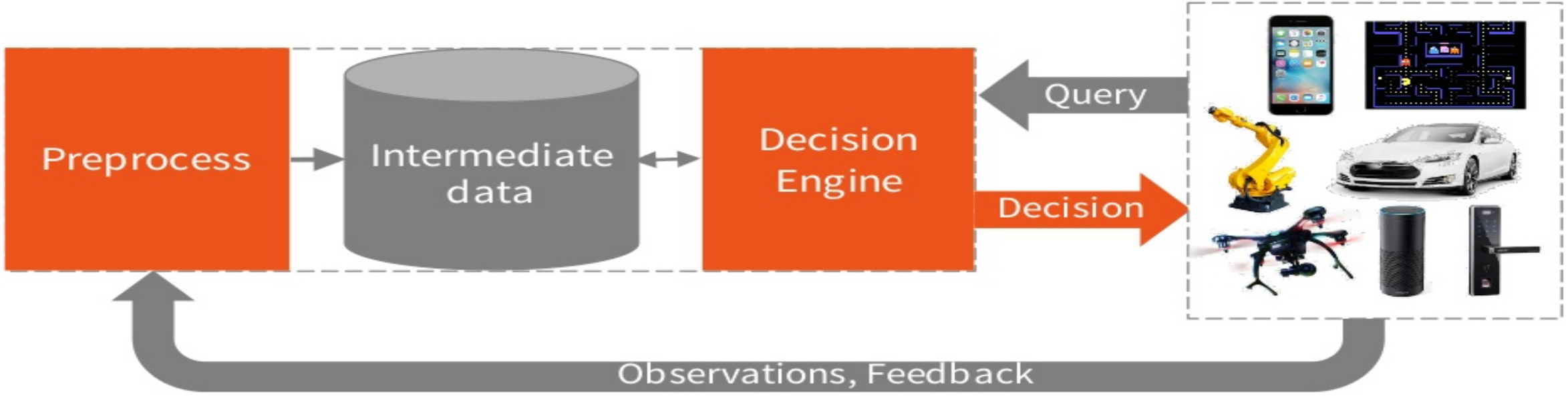
What does this mean?

- Faster decisions better than slower decisions
- Decisions on fresh data better than decisions on stale data
- Decisions on personalized data better than on aggregate data

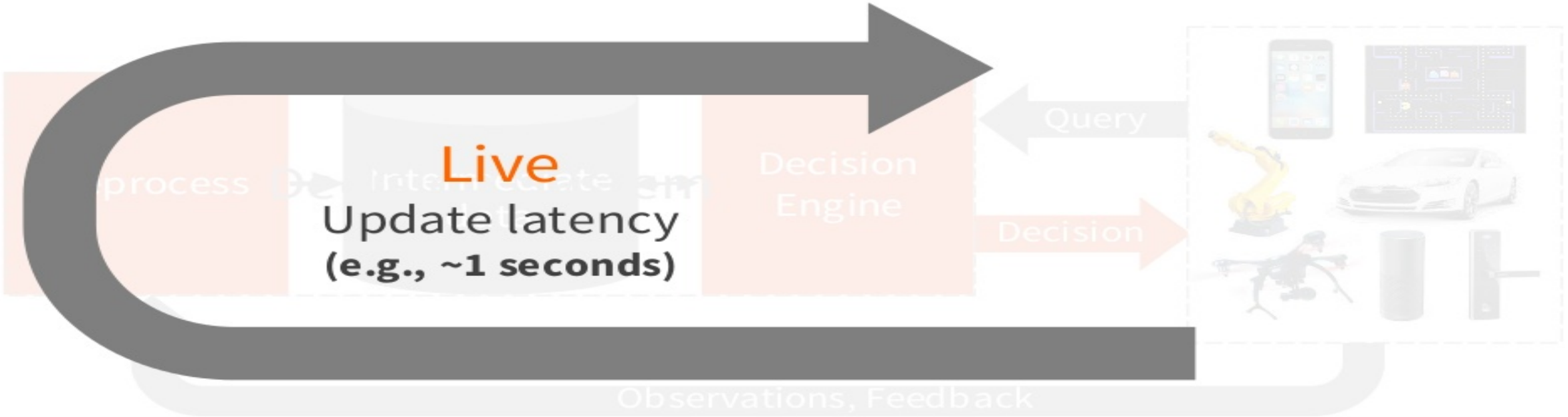
Goal

Real-time decisions	*decide in ms*
on live data	*the current state of the environment*
with strong security	*privacy, confidentiality, integrity*

Typical decision system

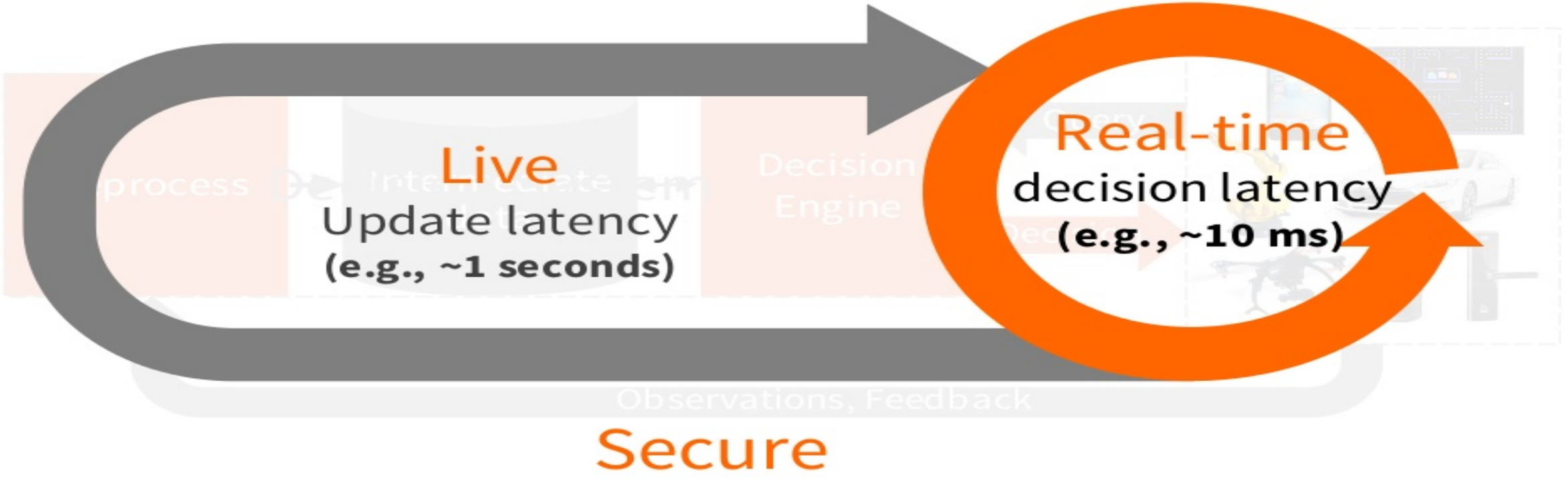


Typical decision system



databricks

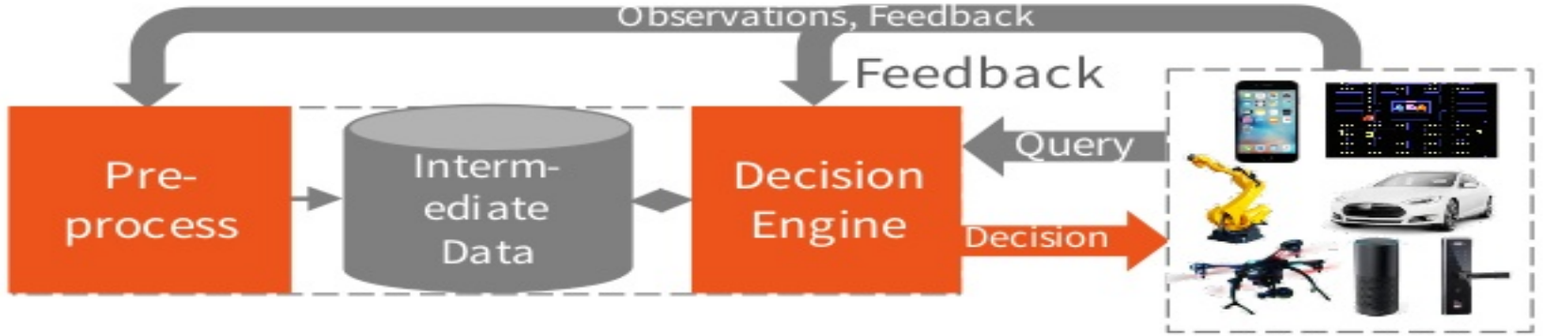
Typical decision system



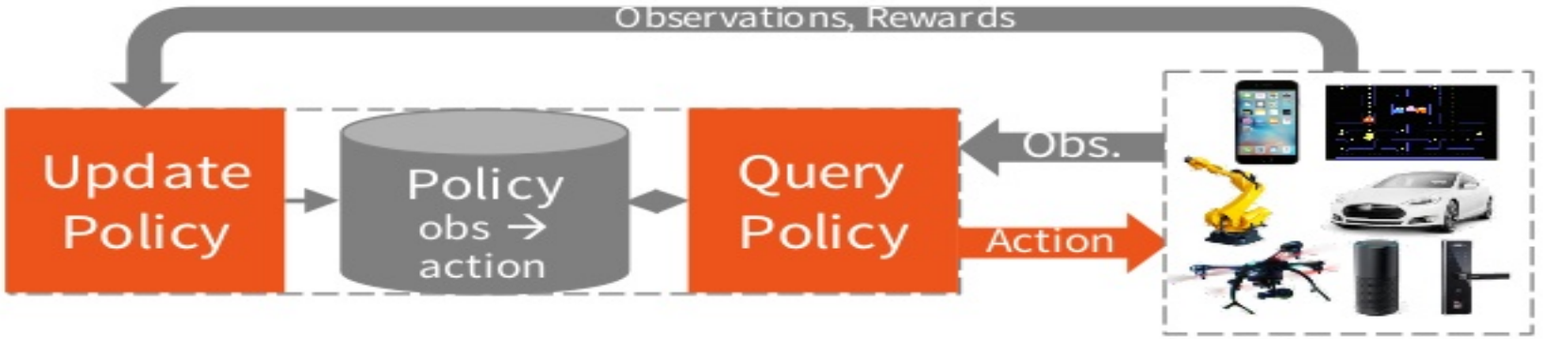
databricks

Example of decision systems

ML Pipeline
(e.g., Clipper + Spark/Tensorflow)



Reinforcement Learning Systems
(e.g., Ray)



What else do we want from decisions?

Intelligent: complex decisions in uncertain environments

Robust: handle complex noise, unforeseen inputs, failures

Explainable: ability to explain non-obvious decisions

Goal

Develop open source
platforms, tools, and algorithms for
intelligent real-time decisions on live-data

Some Proposed Research

Secure Real-time Decisions Stack (SRDS)

- Open source platform to develop of RISE apps
- Secure from ground up
- Reinforcement Learning (RL) as one of key app patterns

Learning control hierarchies: speedup learning, training

Shared learning: learn over confidential data

Some Proposed Research

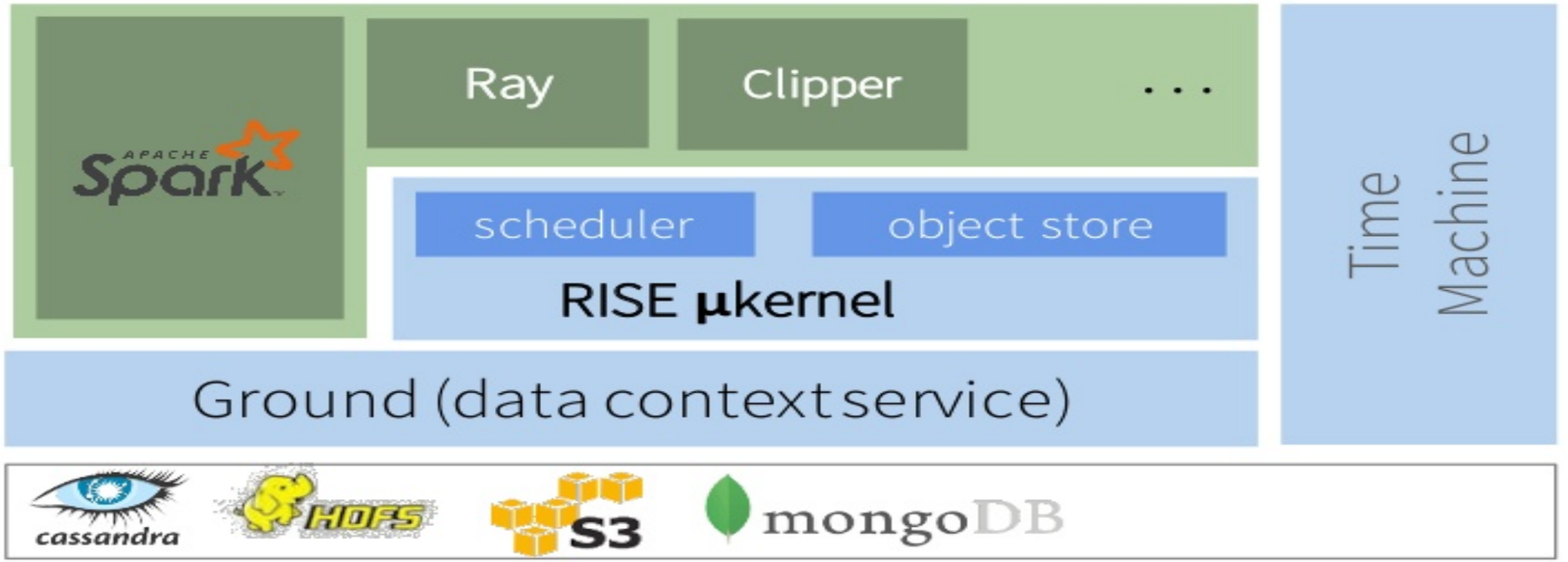
Secure Real-time Decisions Stack (SRDS)

- Open source platform to develop of RISE apps
- Secure from ground up
- Reinforcement Learning (RL) as one of key app patterns

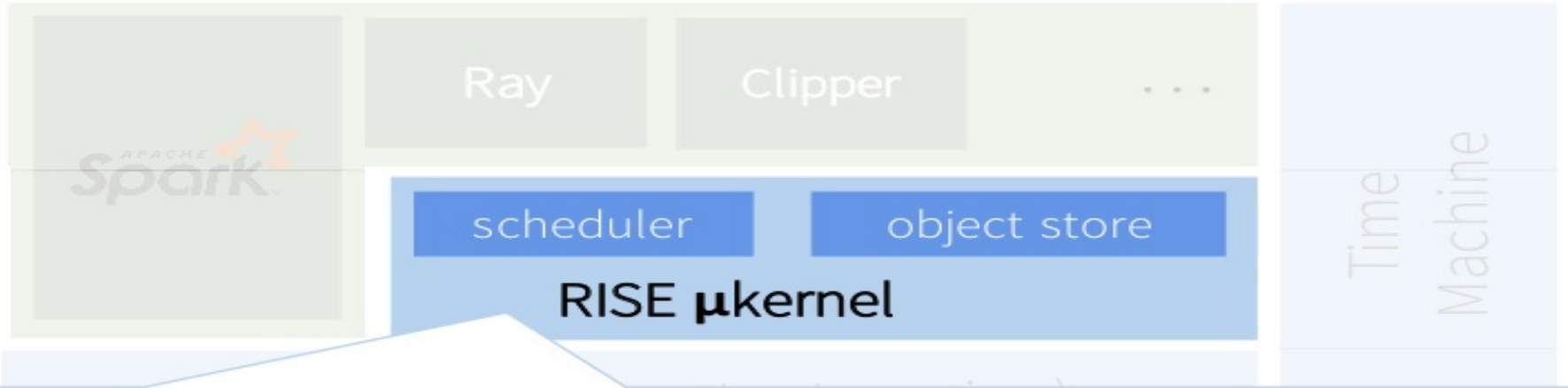
Learning control hierarchies: speedup learning, training

Shared learning: learn over confidential data

Secure Real-time Decision Stack (SRDS)



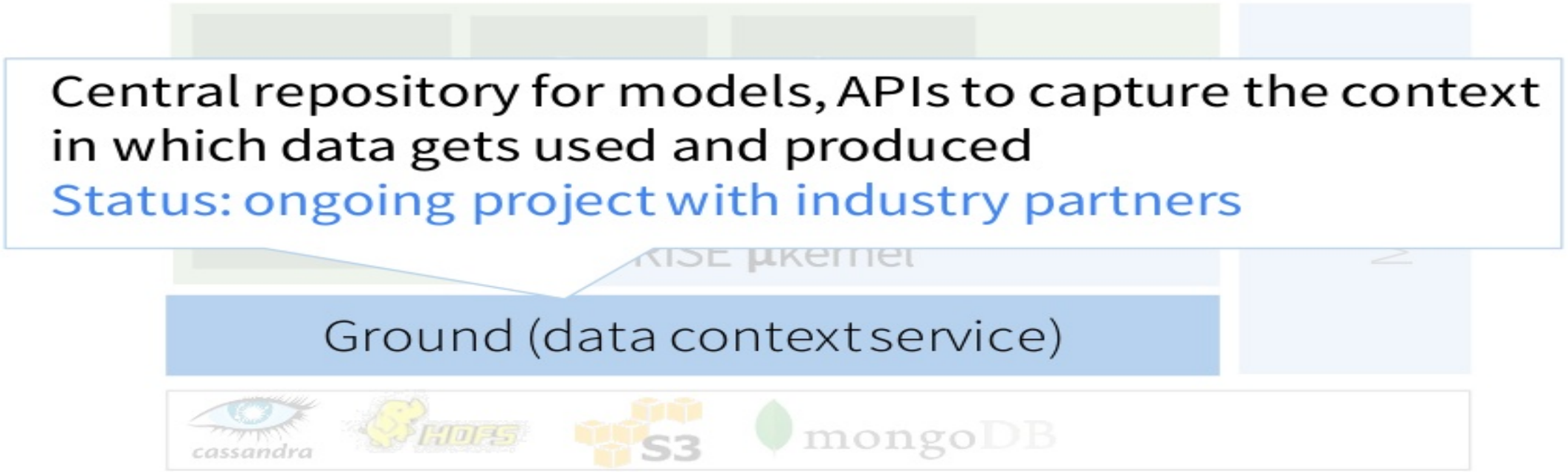
Secure Real-time Decision Stack (SRDS)



Minimalist execution engine:

- Support both data flow and task-parallel execution models
- High-throughput, low-latency: ~ 1M tasks/sec @ ms latency

Secure Real-time Decision Stack (SRDS)



Secure Real-time Decision Stack (SRDS)

Replaying of apps at fine granularity

- Simplify development, debugging
- Robustness: replay against perturbed inputs
- Explainability: identify inputs causing decision
- Security: confirm vulnerabilities, test security patches, compliance auditing

Time Machine

Secure Real-time Decision Stack (SRDS)

Dramatically simplify development of RISE applications

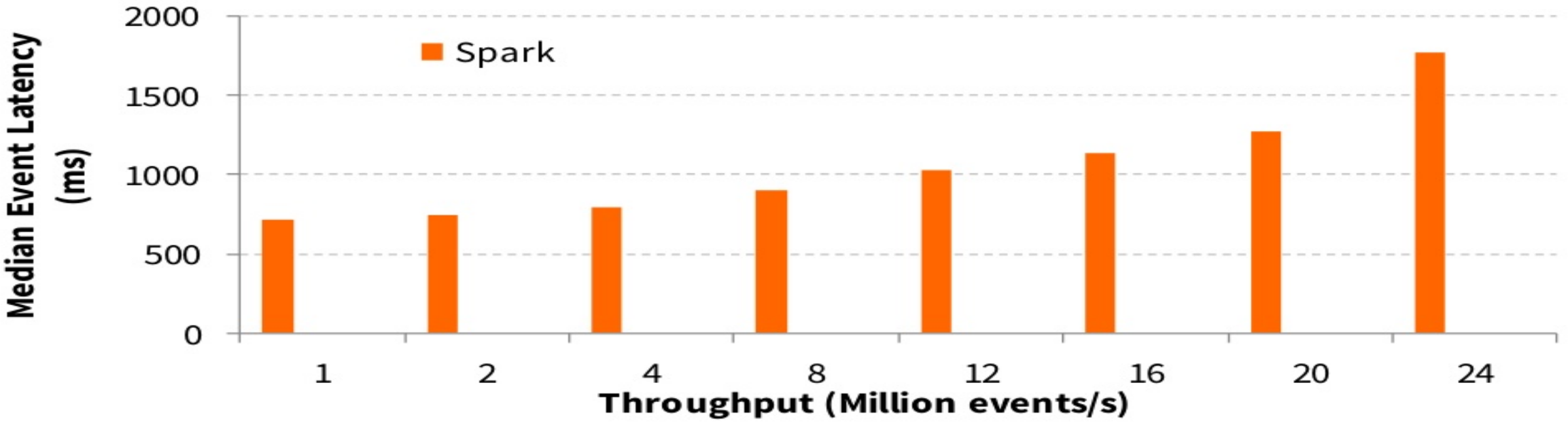
- Apache Spark: improve latency and security
- Clipper: model serving for Apache Spark, Scikit learn, etc
- Ray: framework for RL applications

Improving Apache Spark

Drizzle

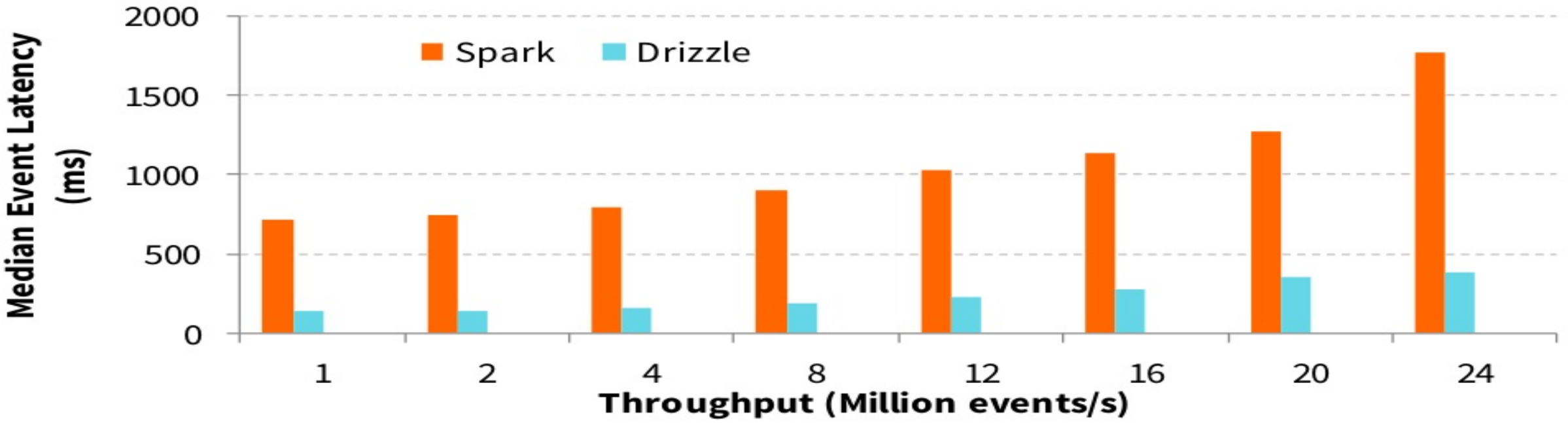
- Decrease latency of Structured Streaming and ML algorithms by ~10x
- Techniques: group scheduling, shared variables

Streaming Latency: YCSB benchmark



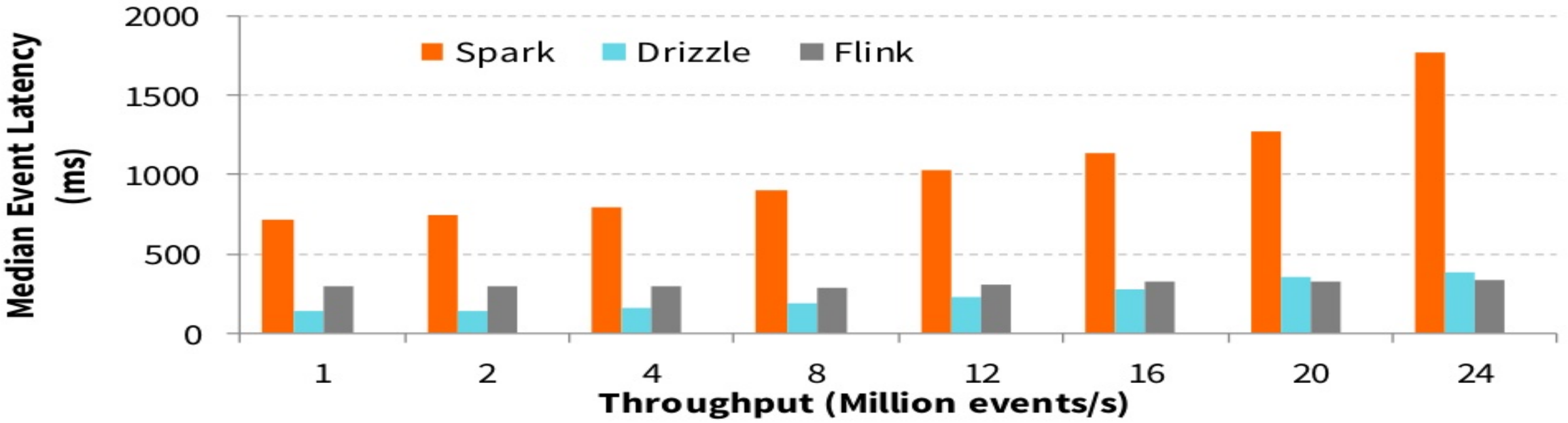
Drizzle-Opt: Reduce-by on mapper side

Streaming Latency: YCSB benchmark



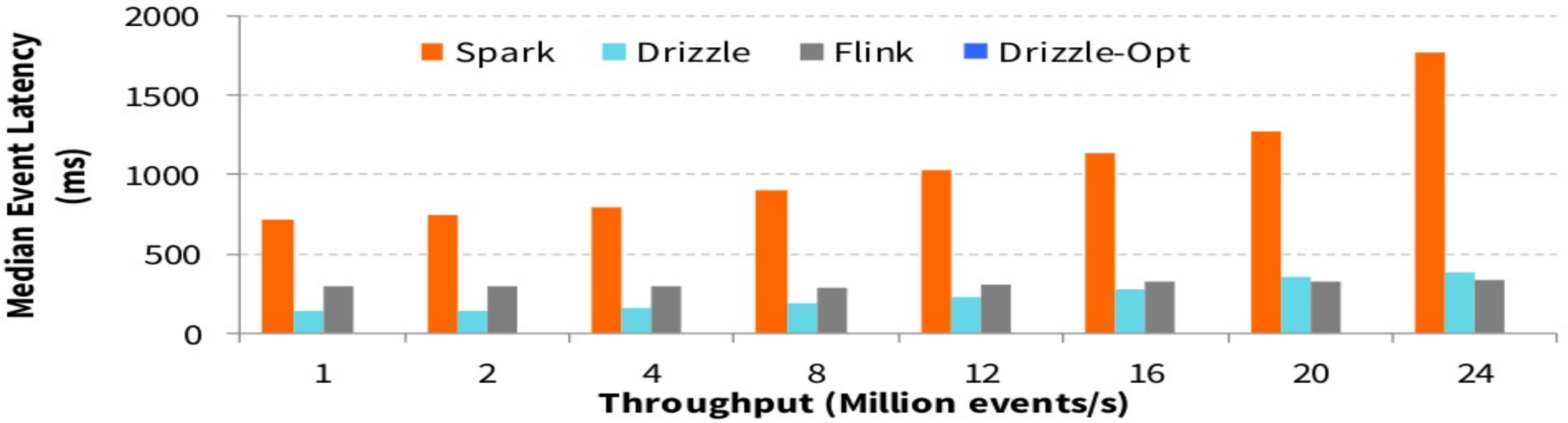
Drizzle-Opt: Reduce-by on mapper side

Streaming Latency: YCSB benchmark



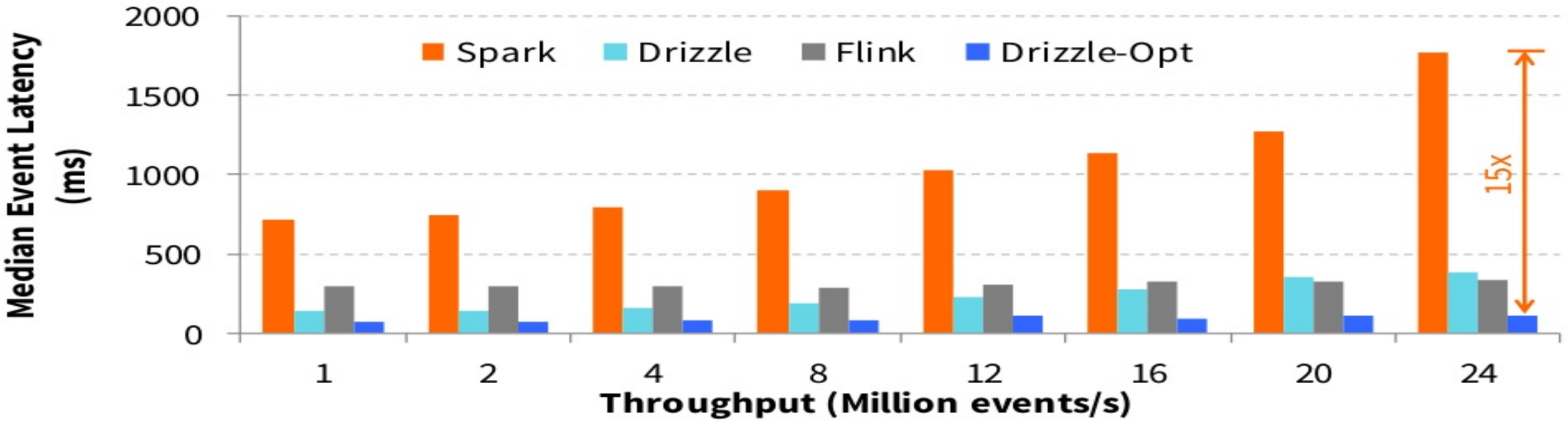
Drizzle-Opt: Reduce-by on mapper side

Streaming Latency: YCSB benchmark



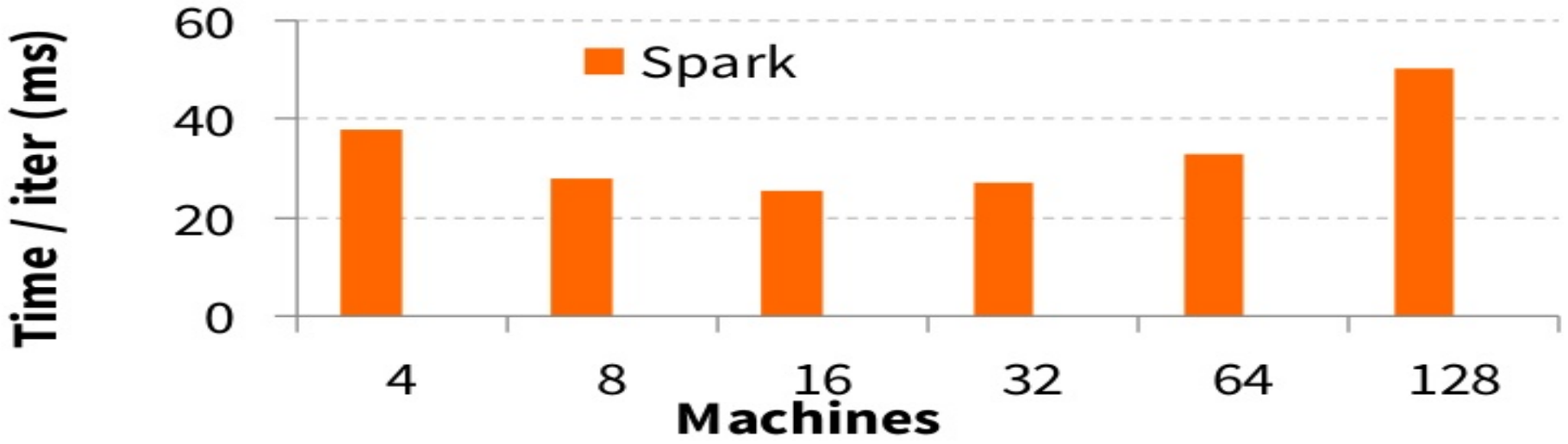
Drizzle-Opt: Reduce-by on mapper side

Streaming Latency: YCSB benchmark

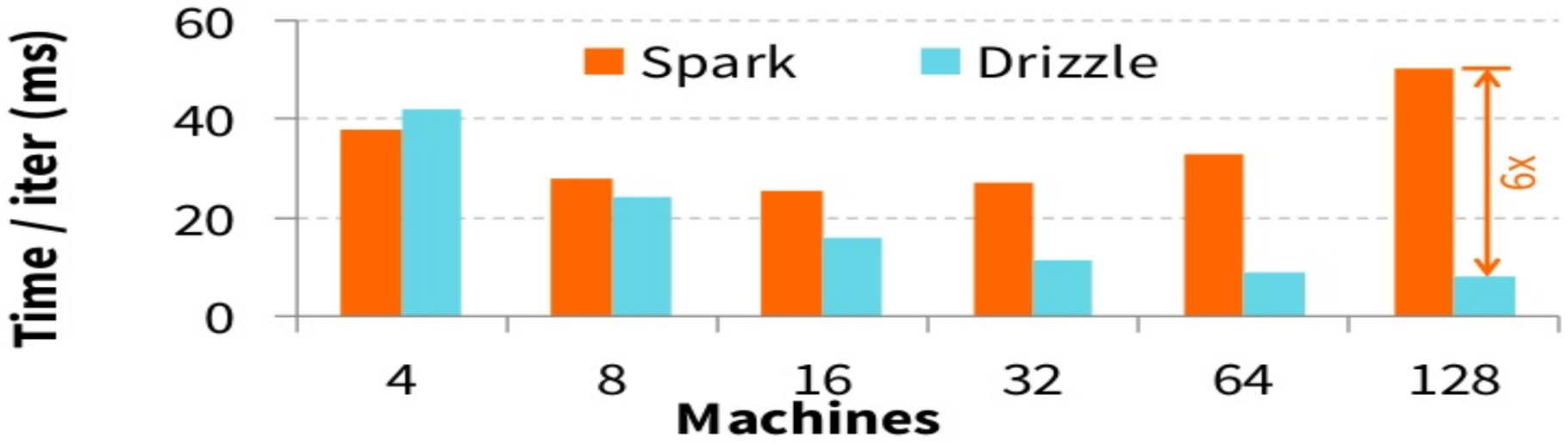


Drizzle-Opt: Reduce-by on mapper side

MLlib: SGD Performance



MLlib: SGD Performance



Improving Apache Spark

Drizzle

- Decrease latency of Structured Streaming and ML algorithms by ~**10x**
- Techniques: group scheduling, shared variables
- Some of these techniques will make their way to Apache Spark

Opaque

- Full data encryption, authentication, and verification (Intel's SGX)
- Oblivious mode: hide data access pattern
- Support most SparkSQL functionality
- See Wenting's talk later

RISELab

Goal: Develop open source platforms, tools, and algorithms for intelligent real-time decisions on live-data

Already promising results

Expect much more over the next five years!

Thank you