Apache crail

Accelerating workloads with a high-performance distributed ephemeral data store

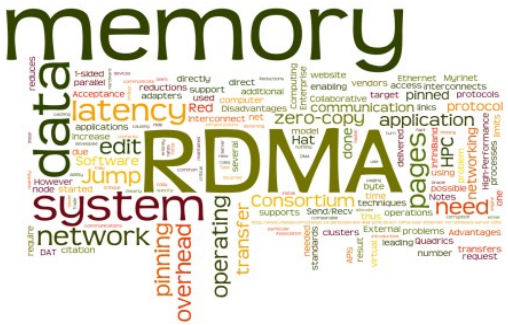
-

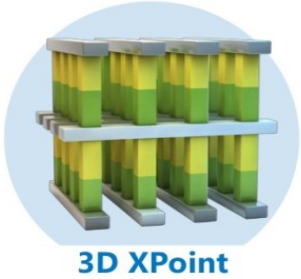
Patrick Stuedi & Jonas Pfefferle
IBM Research

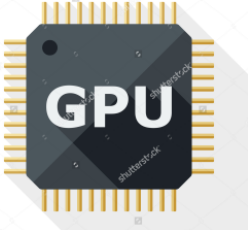
Trend1: I/O Performance is increasing dramatically

2008		**2018+**
100MB/s 10ms	100x →	10GB/s 10µs
1Gbps 100µs		200Gbps 1µs

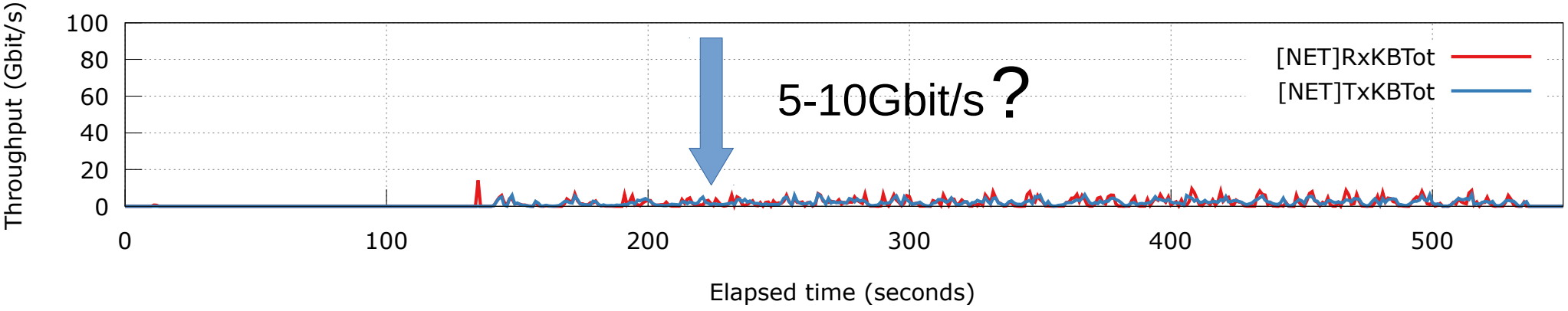
Trend2: I/O diversity is increasing dramatically



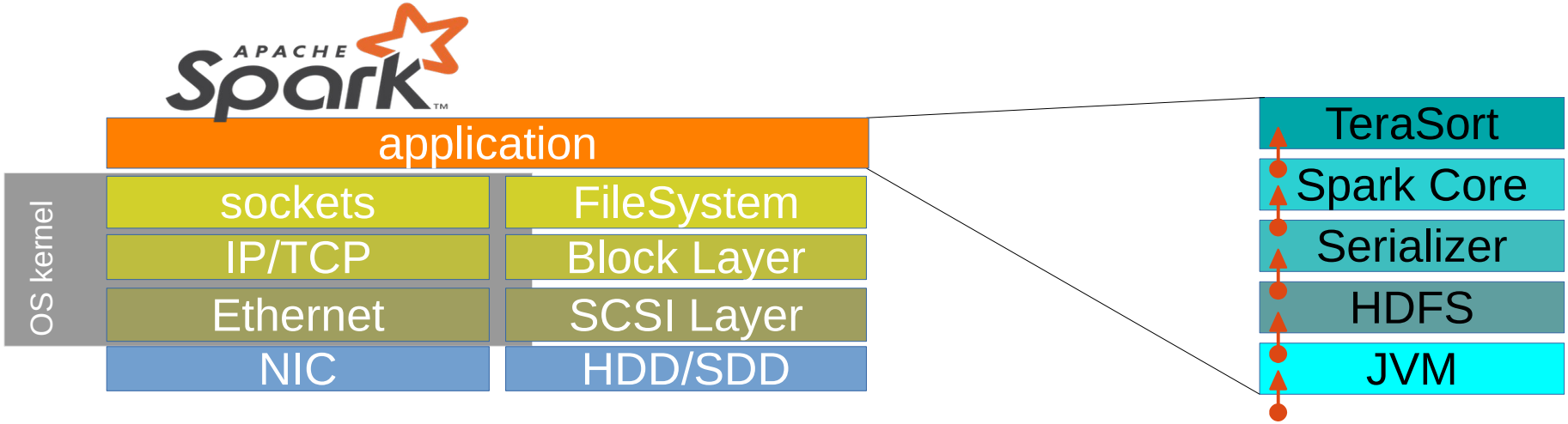




Spark TeraSort @100GbE



What is the Problem



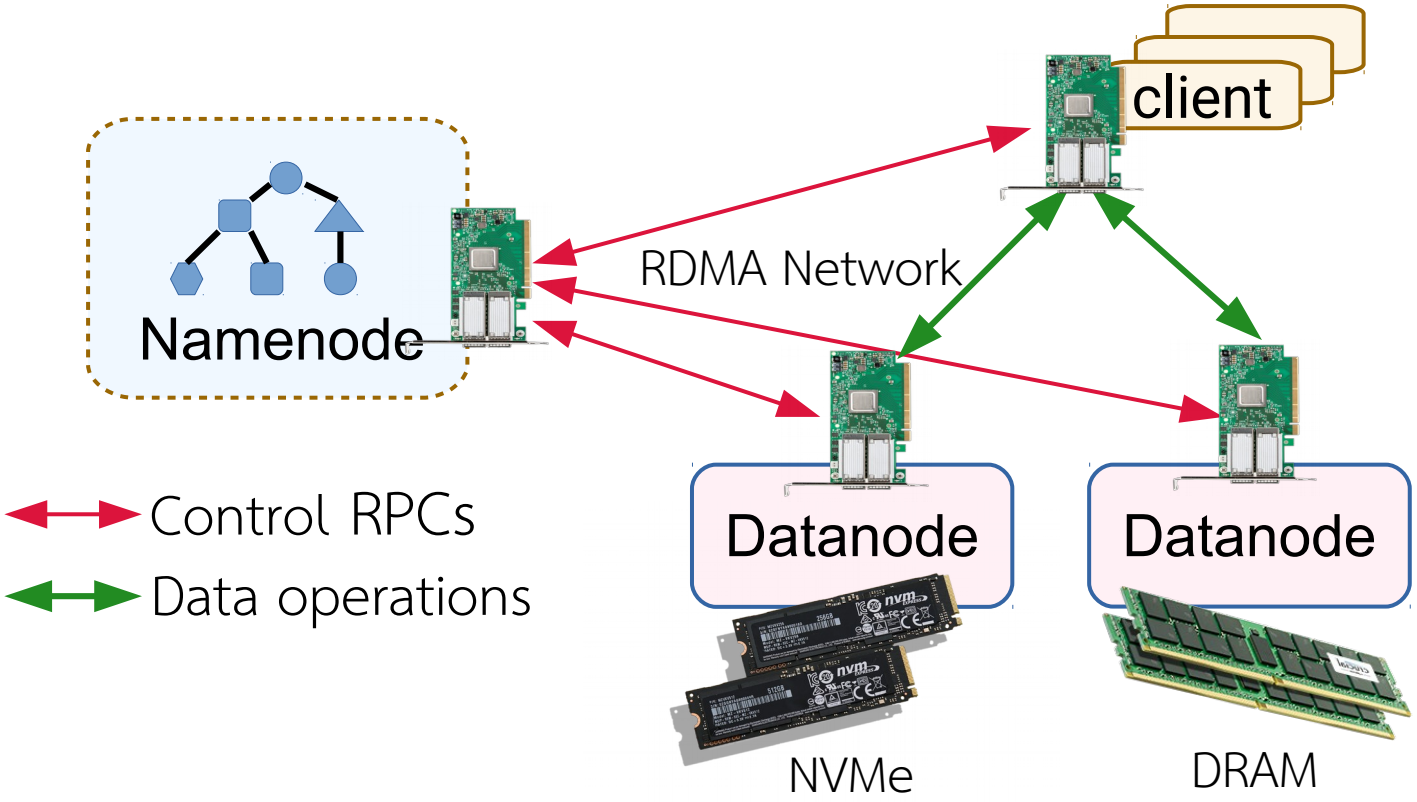
- Application use the legacy APIs
- Applications themselves are heavily layered!
- Overhead during local file system writing
- Overhead during network processing

Challenge:
Leverage the speed of I/O hardware.

Apache crail

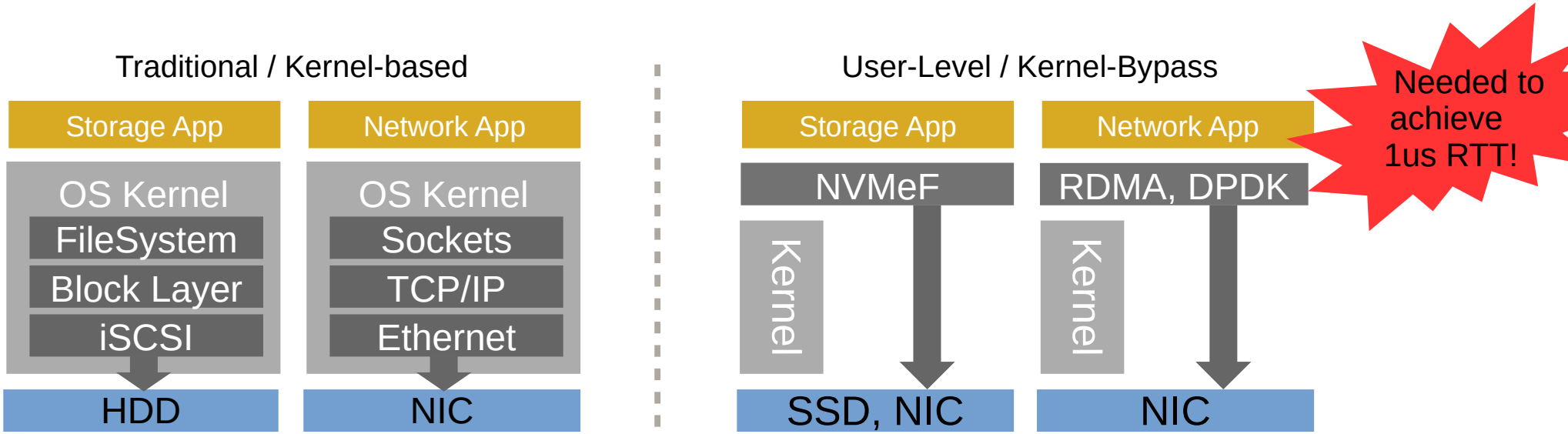
- A novel I/O architecture for data processing systems
- Distributed data store for ephemeral data
- Designed for new I/O hardware (RDMA, NVMe, …)
- Speed-up workloads by accelerating data sharing
- Apache incubator project => Open source!

System Overview



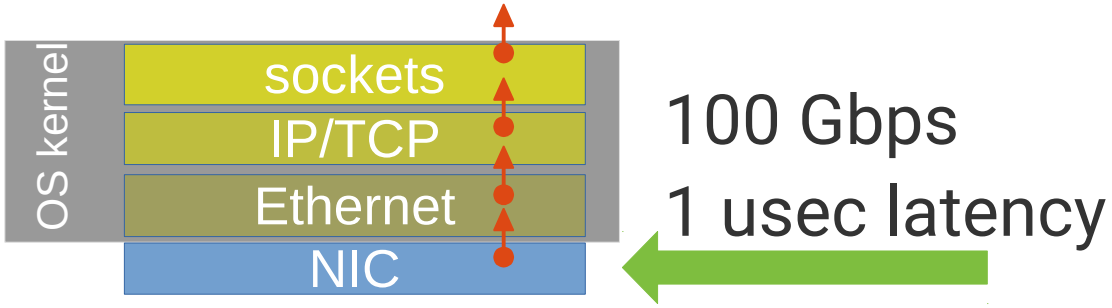
Challenge:
Deliver I/O performance to the Application.

Design Principles – I

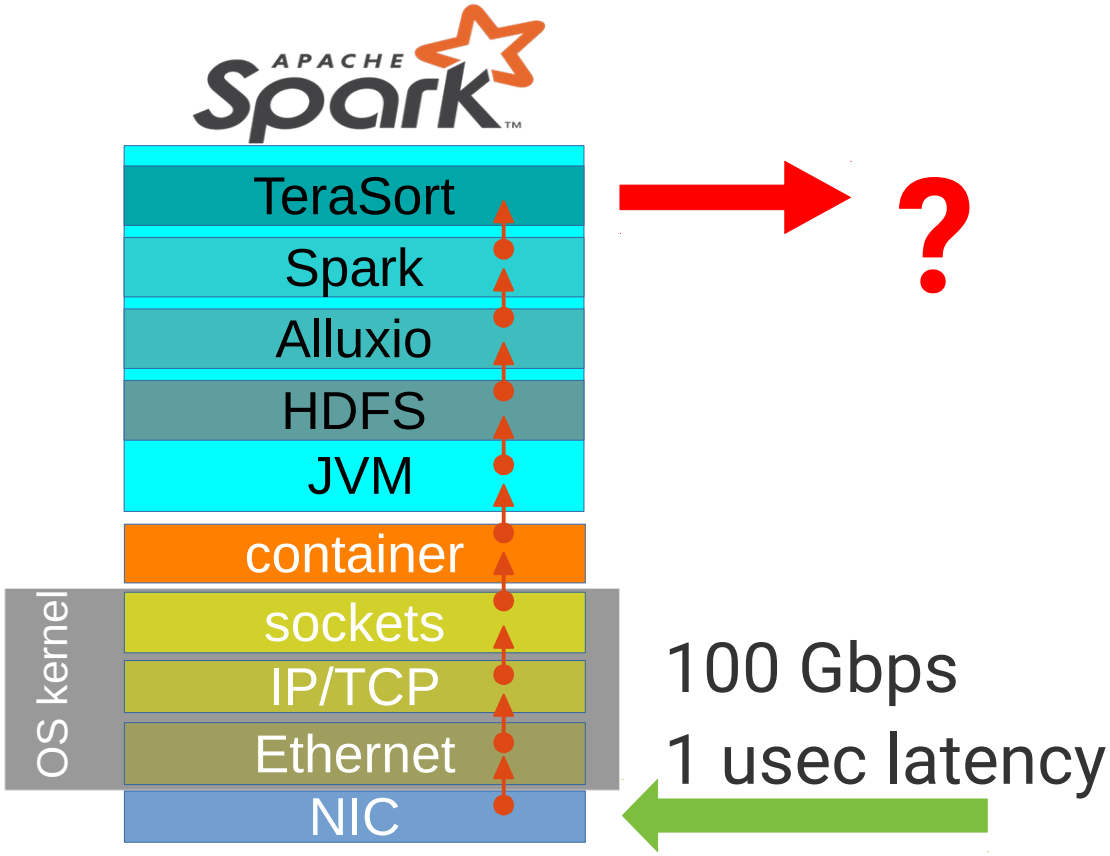


Modern APIs for Networking and Storage offer asynchronous non-blocking user-level access to hardware

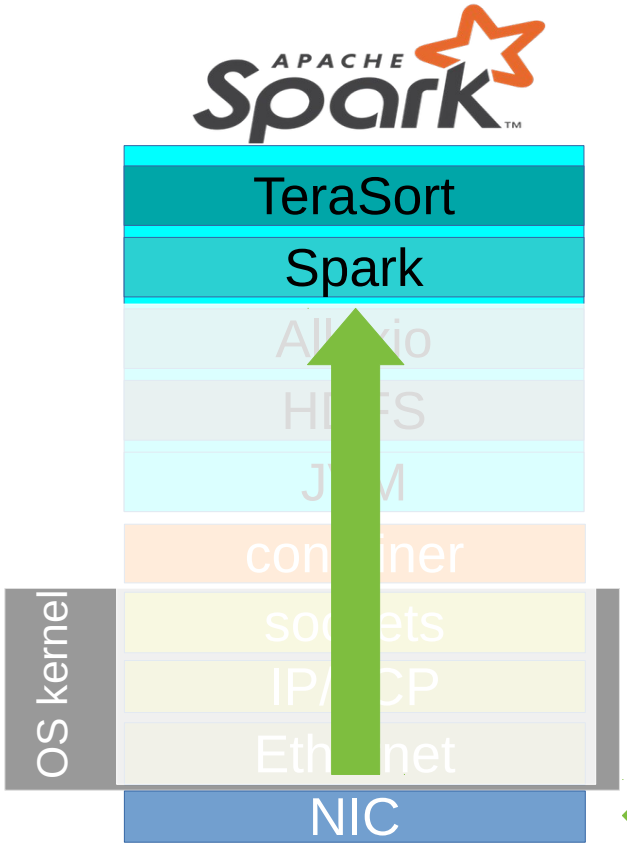
Design Principles – I



Design Principles – I



Design Principles – I



1. Use high-performance user-level I/O

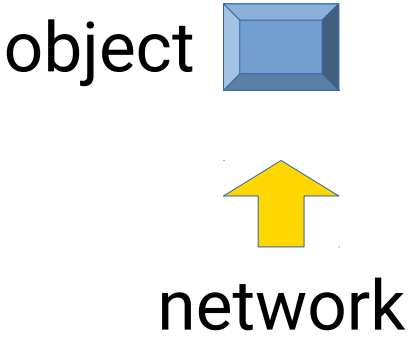
RDMA[1], NVMeF[2], SPDK/DPDK in Java

- one-sided RDMA read/write operations

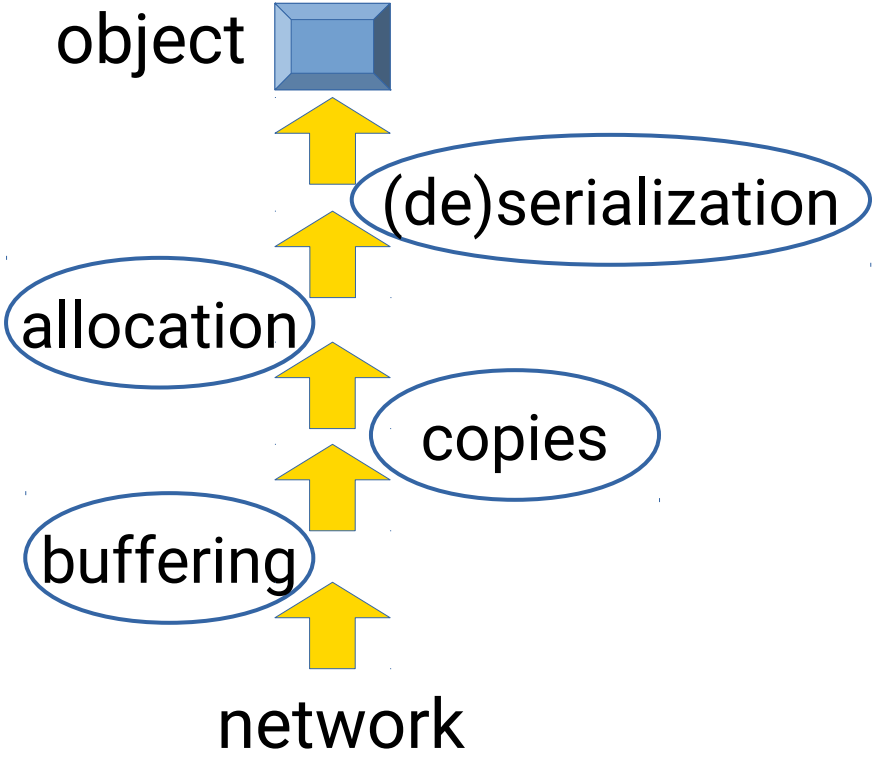
100 Gbps

1 usec latency

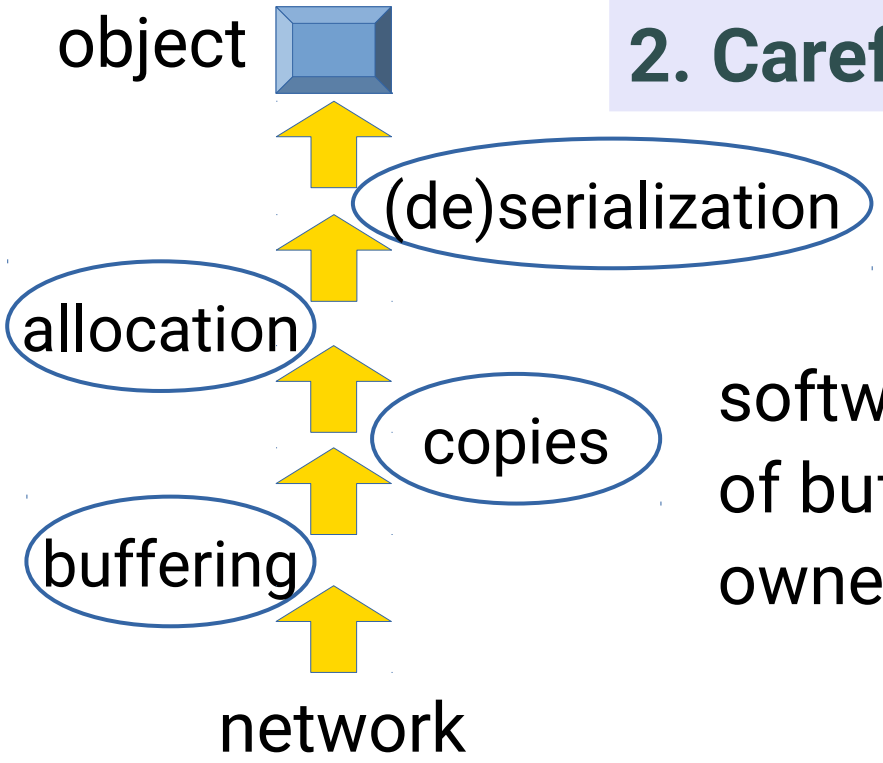
Design Principles – II



Design Principles – II



Design Principles – II

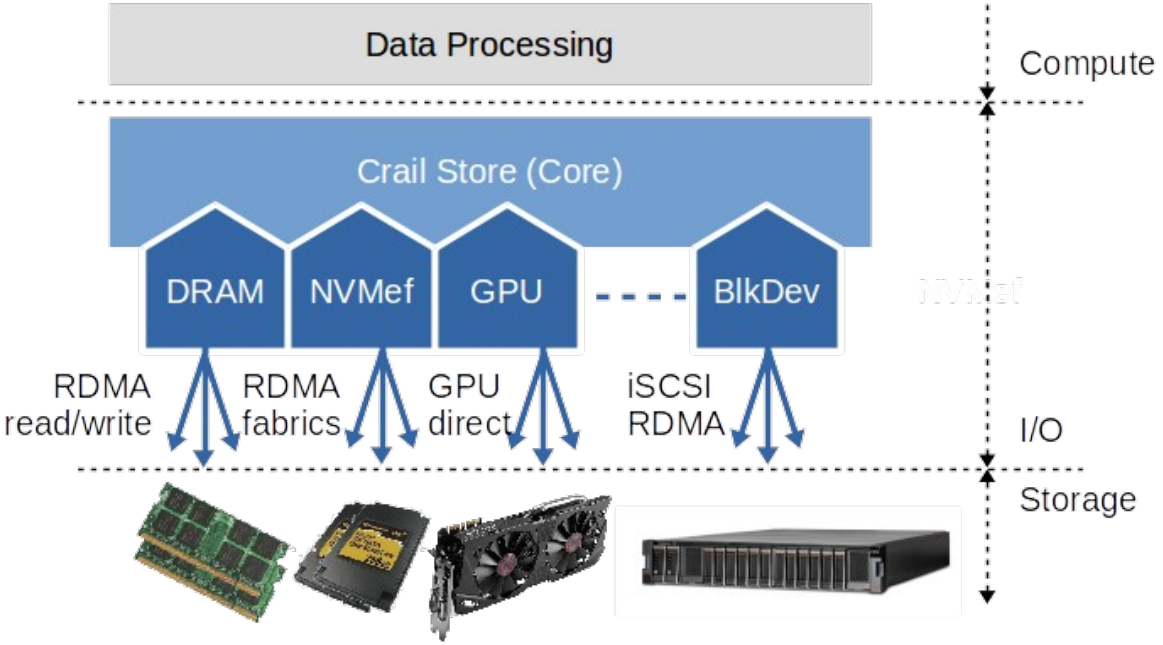


2. Careful software design for the µsec-Era

software time budgeting, pre-allocation of buffers, buffer caching, and reusing, ownership, etc.

Design Principles – III

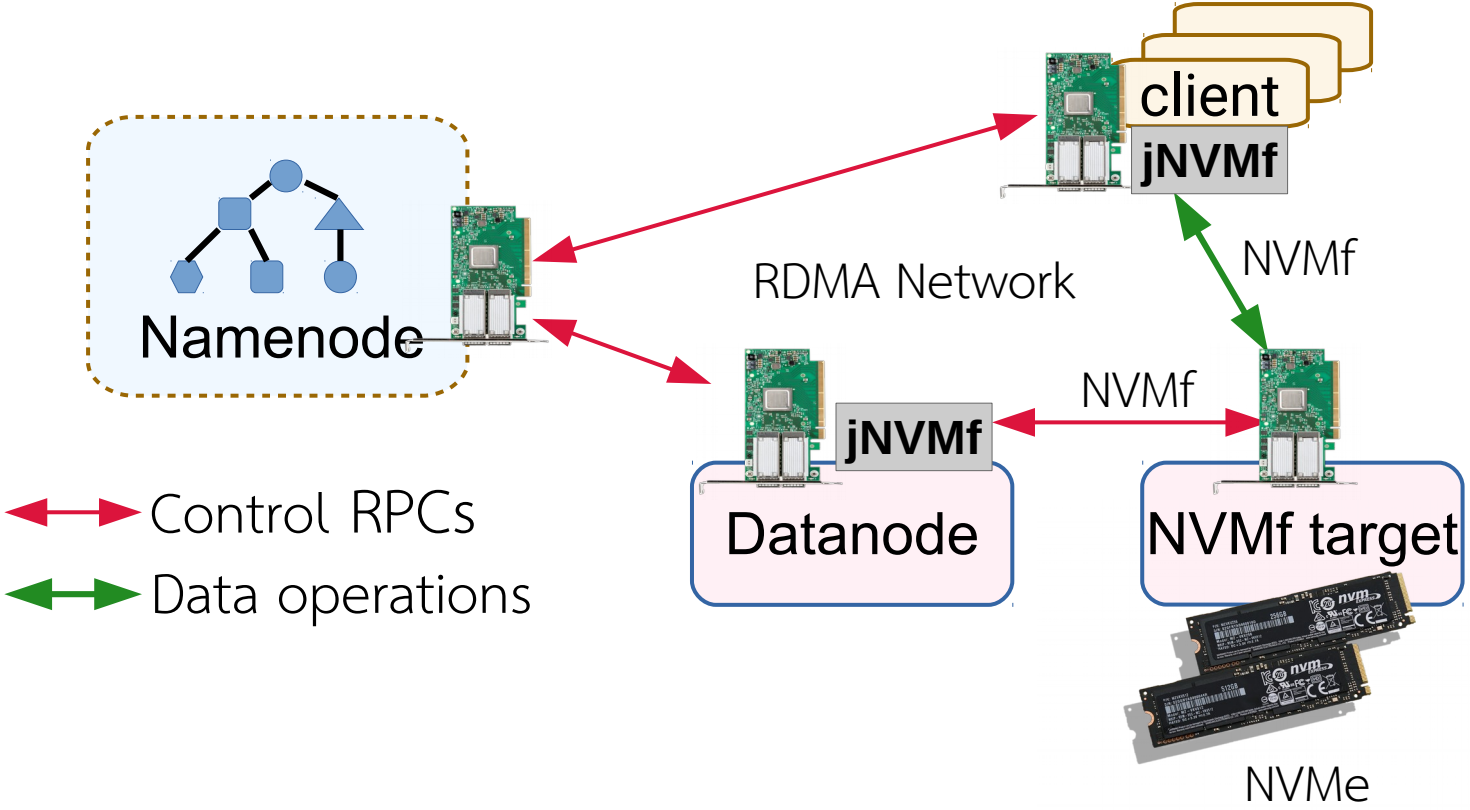
3. Support heterogeneous I/O hardware



NVMf Storagetier

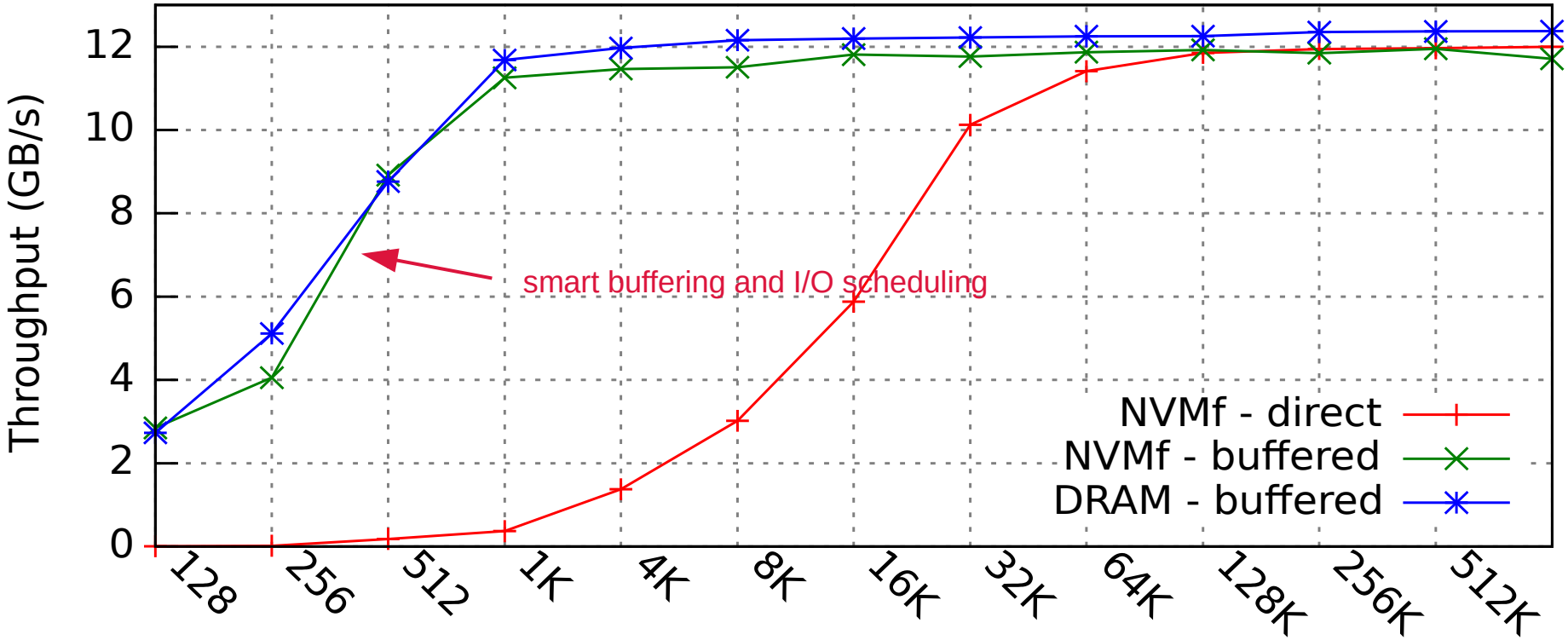
- NVMe over Fabrics
  - Access to remote storage
  - NVMe commands over RDMA
- One known user-level implementation: SPDK
  - No Java support
  - Started with API wrapper for Java => not a good fit
- Own user-level initiator implementation: **jNVMf**
  - Standard compliant: connects to any target
  - Supports advanced NVMe/f features like inline data, arbitration, etc.
  - *https://github.com/zrlio/jNVMf*

NVMf Storagetier



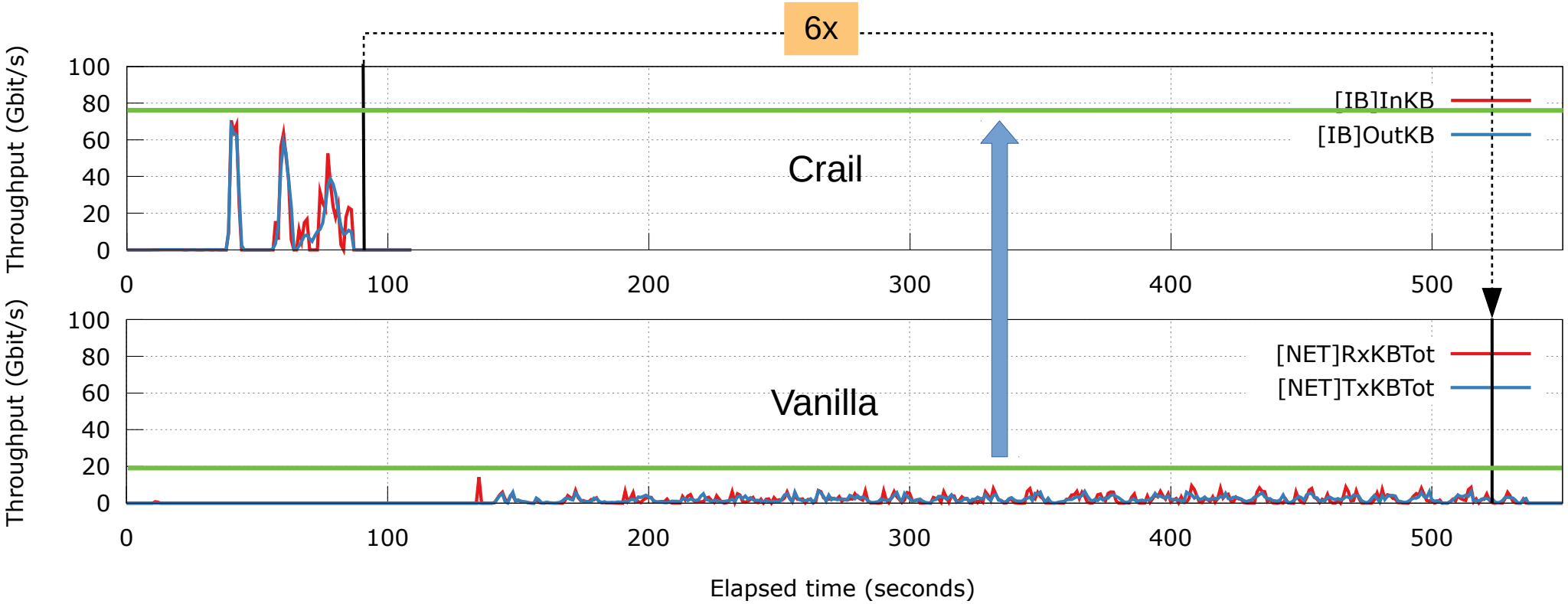
Performance Evaluation

NVMf Read Throughput

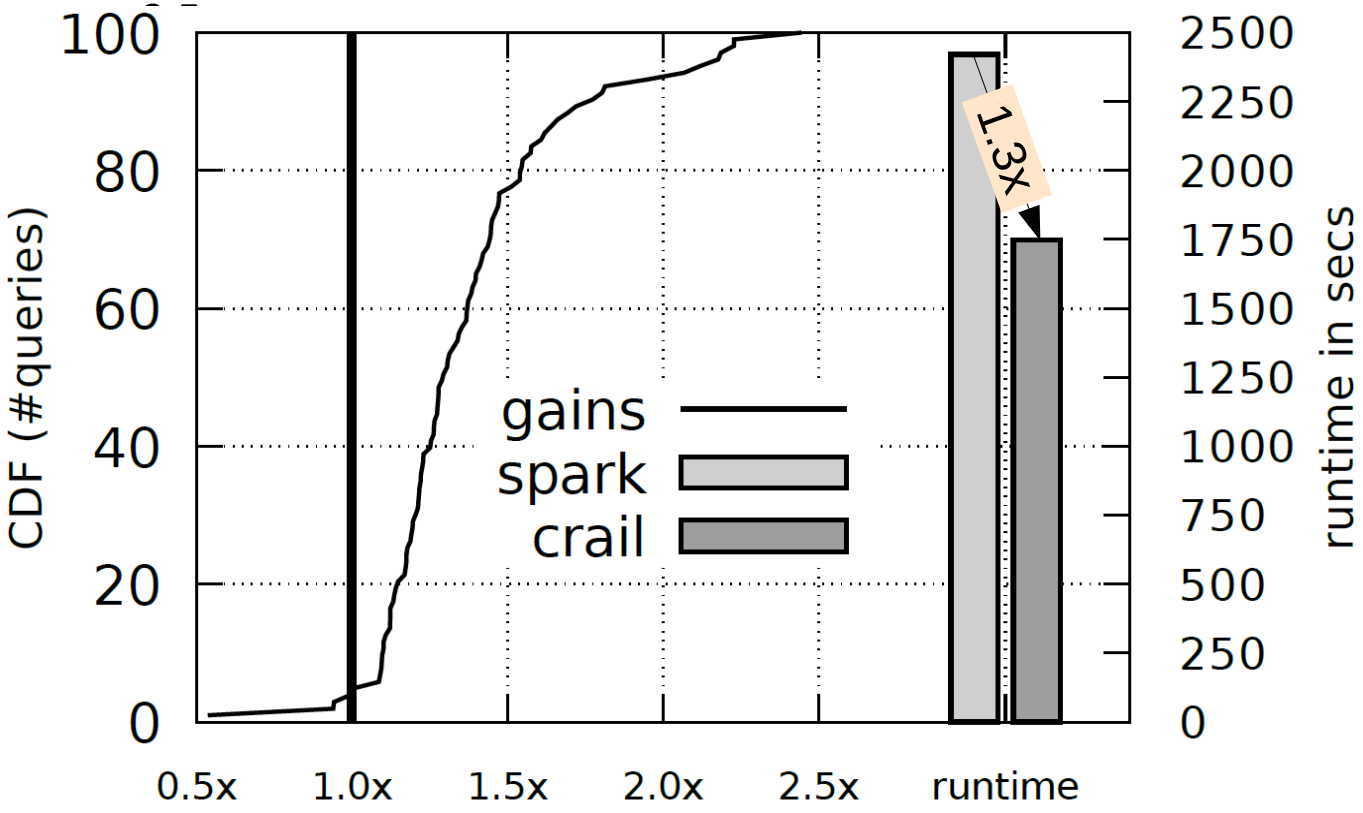


Crail delivers full hardware bandwidth from remote NVMe flash

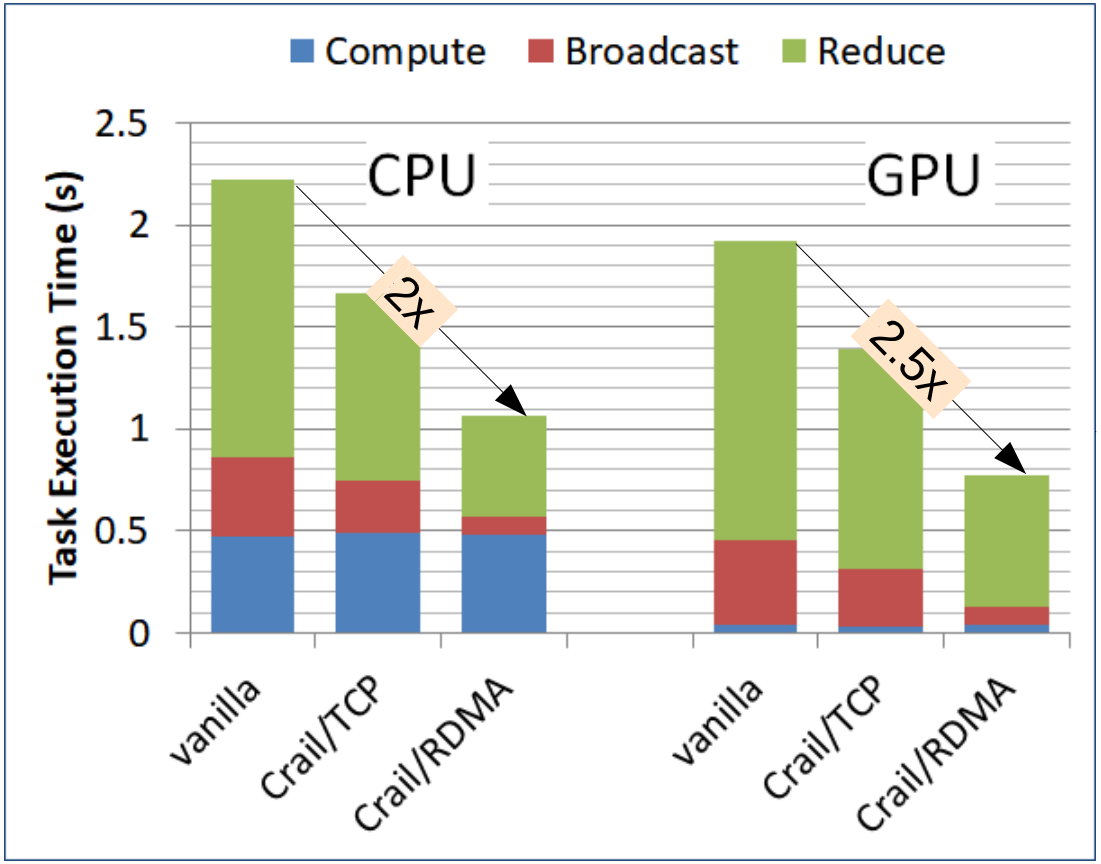
Spark TeraSort



Spark/SQL - TPC-DS



Spark/SnapML – Logistic Regression



Thanks!

crail.apache.org

Contact us:
dev@crail.apache.org