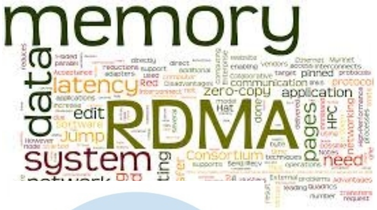
SPARK SUMMIT 2017

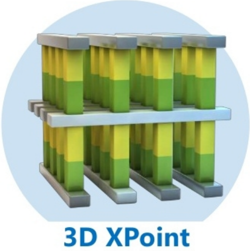
Running Spark on a High-Performance Cluster using RDMA Networking and NVMe Flash

Patrick Stuedi, IBM Research

Hardware Trends

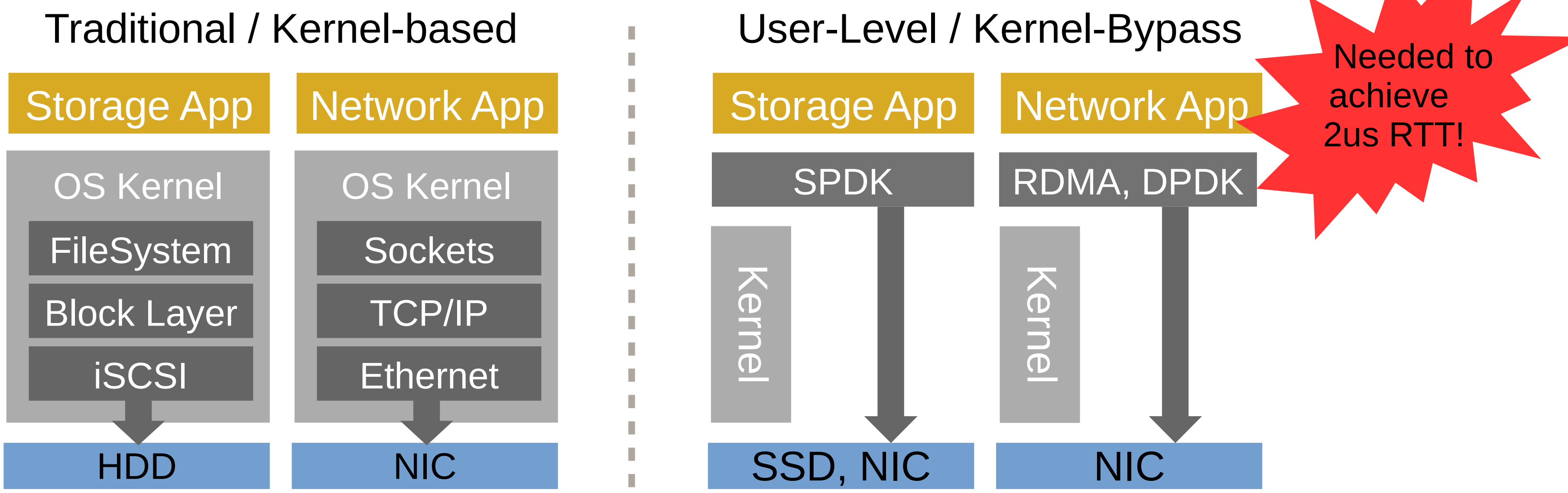
		community target	our target
	2010	2017	2017
Storage	100 MB/s 100ms	1000 MB/s 200us	10 GB/s 50us
Network	1Gbps 50us	10Gbps 20us	100Gbps 2us
CPU	~3GHz	~3GHz	☹





nvm EXPRESS®

User-Level APIs



Traditional / Kernel-based

Storage App

OS Kernel

FileSystem

Block Layer

iSCSI

HDD

Network App

OS Kernel

Sockets

TCP/IP

Ethernet

NIC

User-Level / Kernel-Bypass

Storage App

SPDK

Kernel

SSD, NIC

Network App

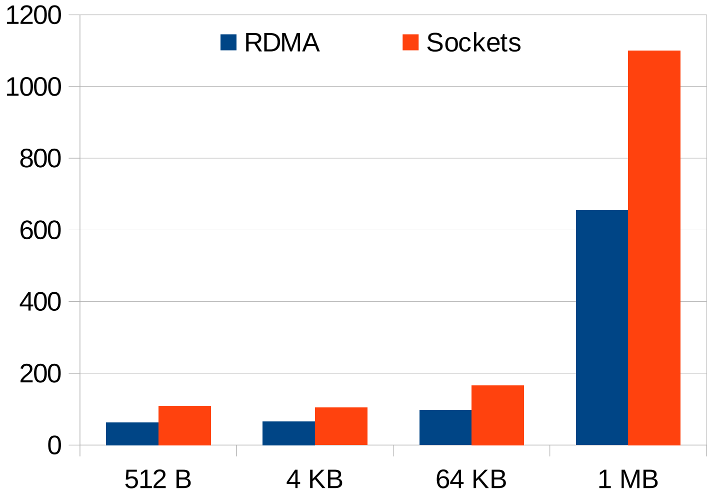
RDMA, DPDK

Kernel

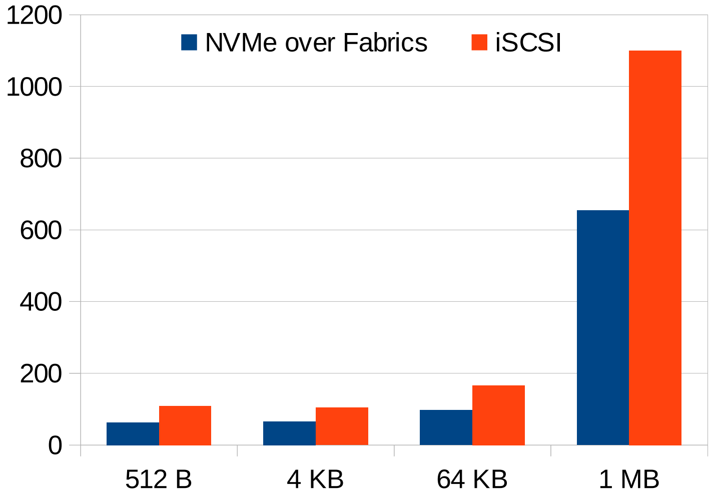
NIC

Needed to achieve 2us RTT!

Remote Data Access

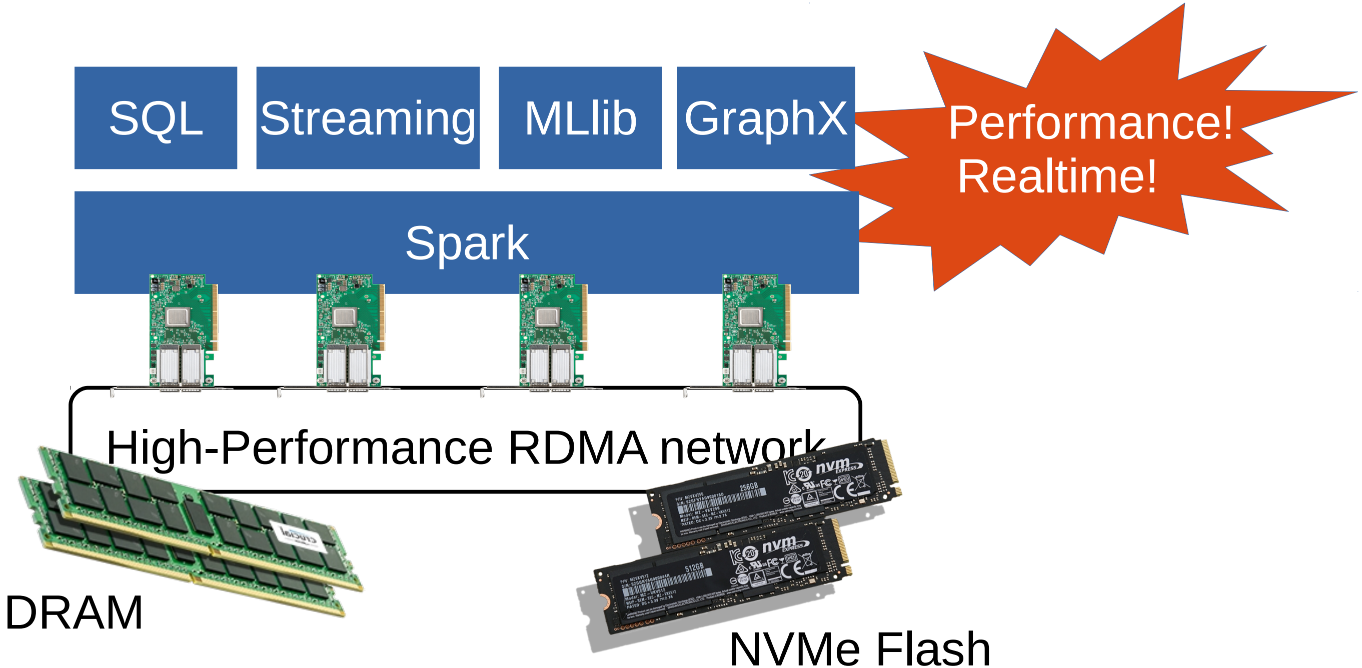


DRAM



NVMe Flash

Let's Use it!



SQL

Streaming

MLlib

GraphX

Spark

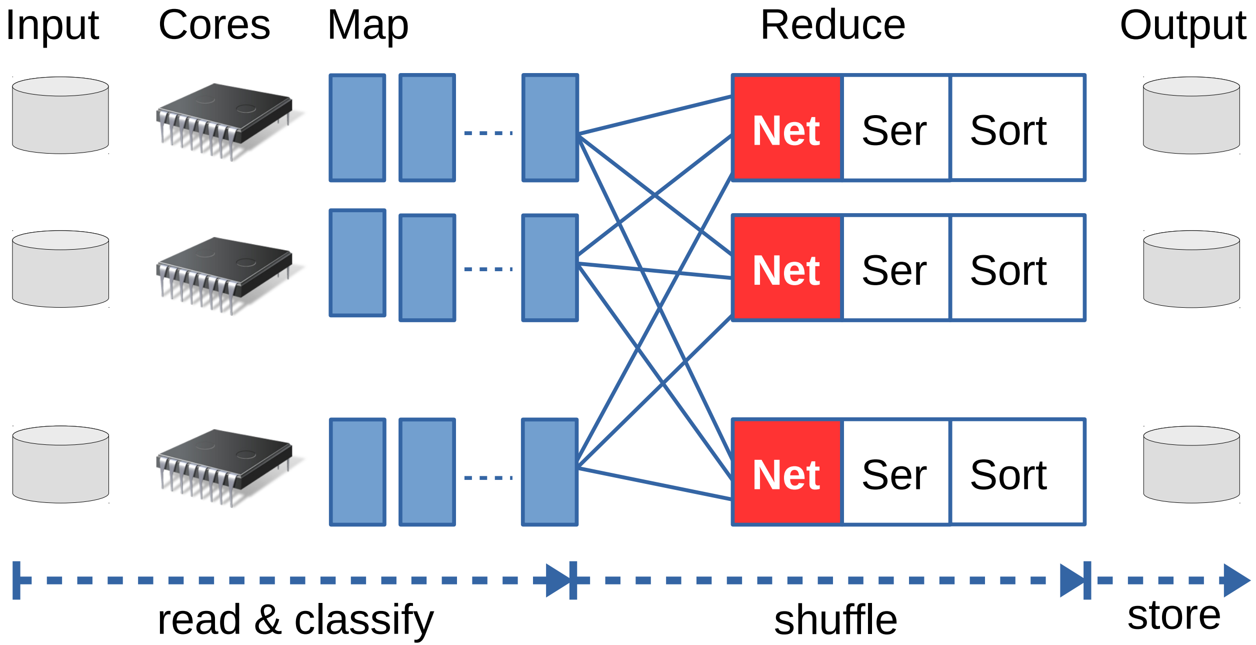
Performance!
Realtime!

High-Performance RDMA network

DRAM

NVMe Flash

Case Study: Sorting in Spark

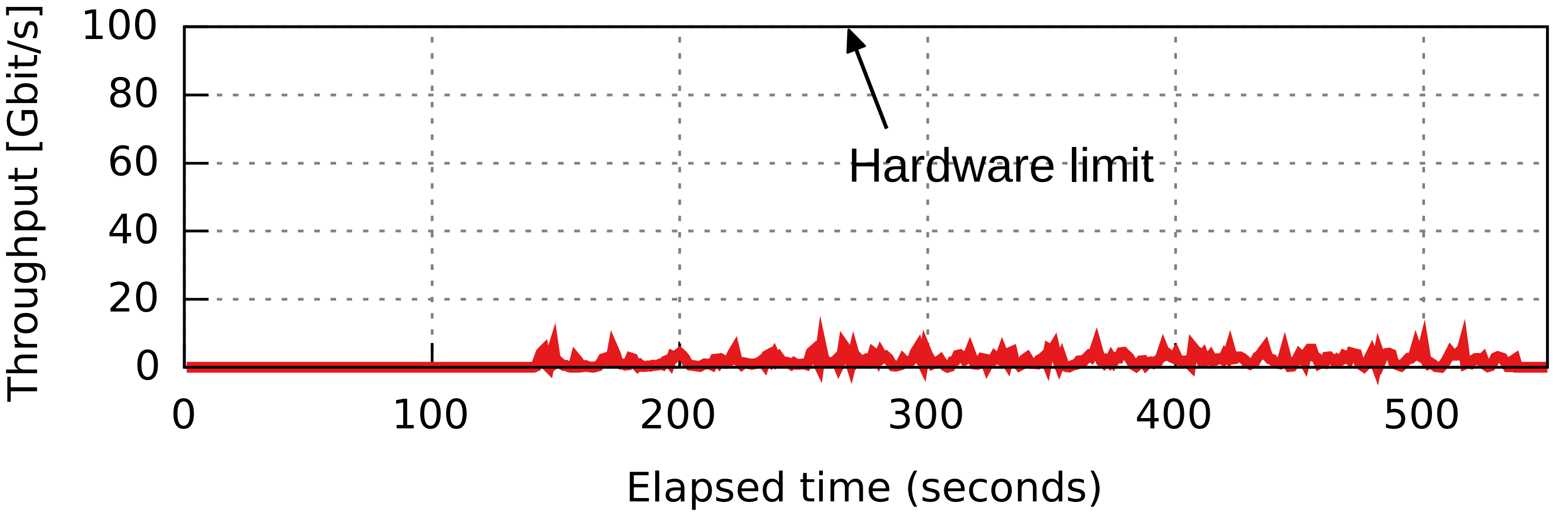


Experiment Setup

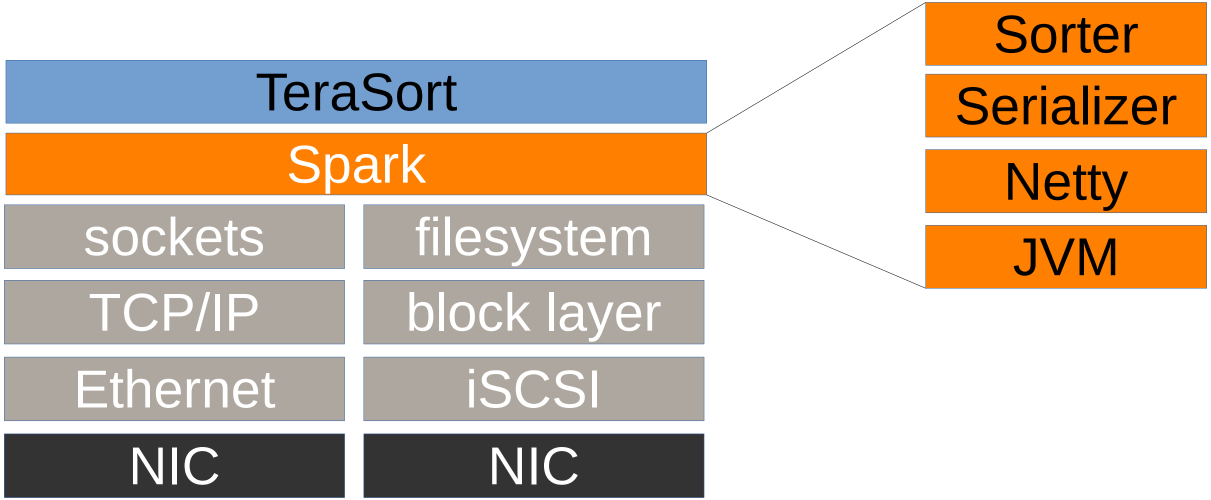
- Total data size: 12.8 TB
- Cluster size: 128 nodes
- Cluster hardware:
  - DRAM: 512 GB DDR 4
  - Storage: 4x 1.2 TB NVMe SSD
  - Network: 100GbE Mellanox RDMA

Flash bandwidth per node matches network bandwidth

How is the Network Used?

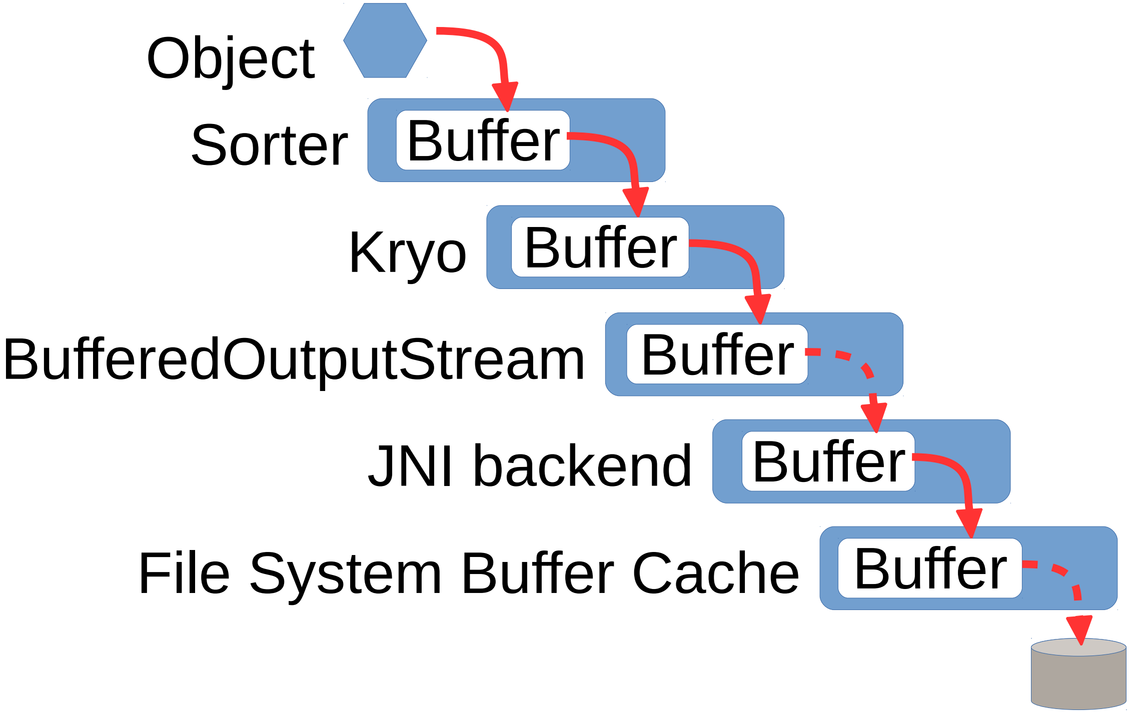


What are the Problems?

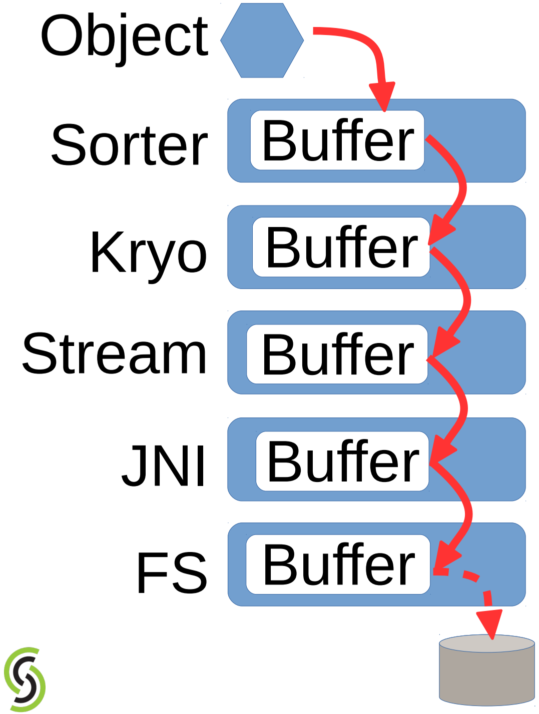


- Spark uses legacy networking and storage APIs: no kernel-bypass
- Spark itself introduces additional I/O layers: Netty, serializer, sorter, etc.

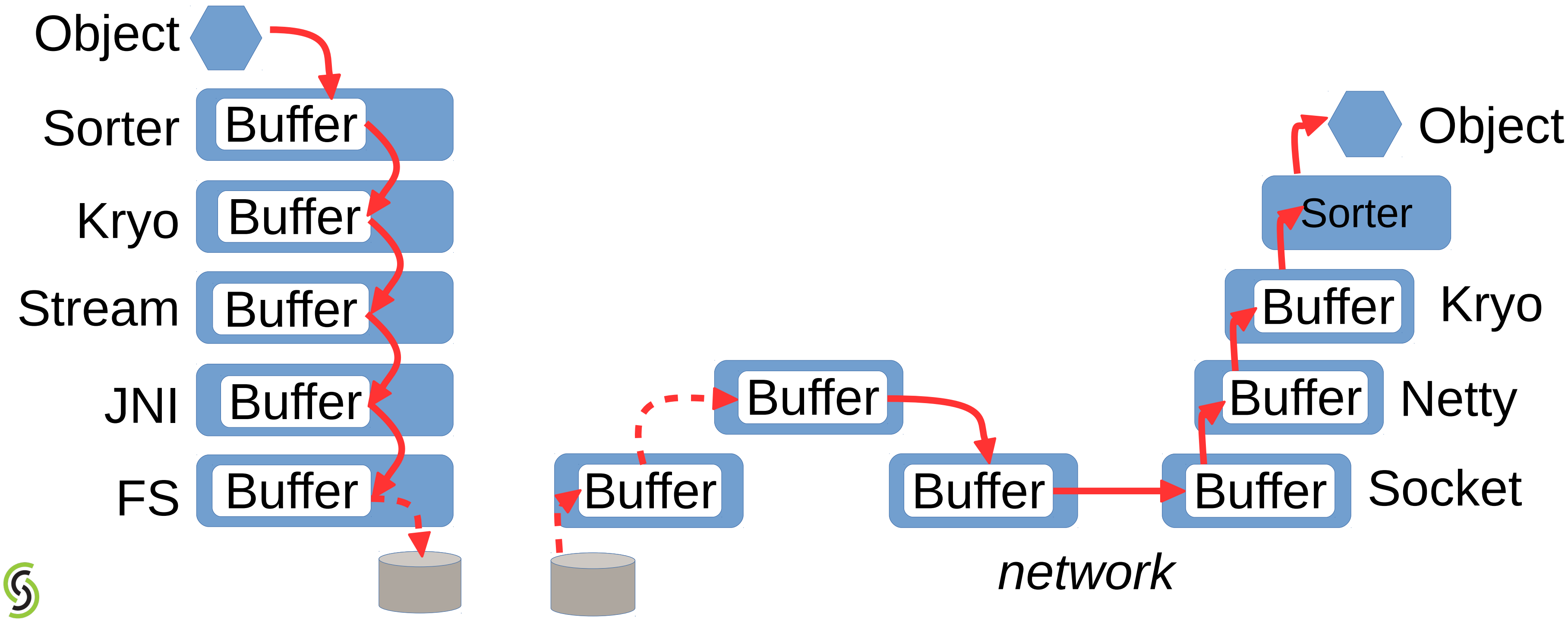
Example: Shuffle (Map)



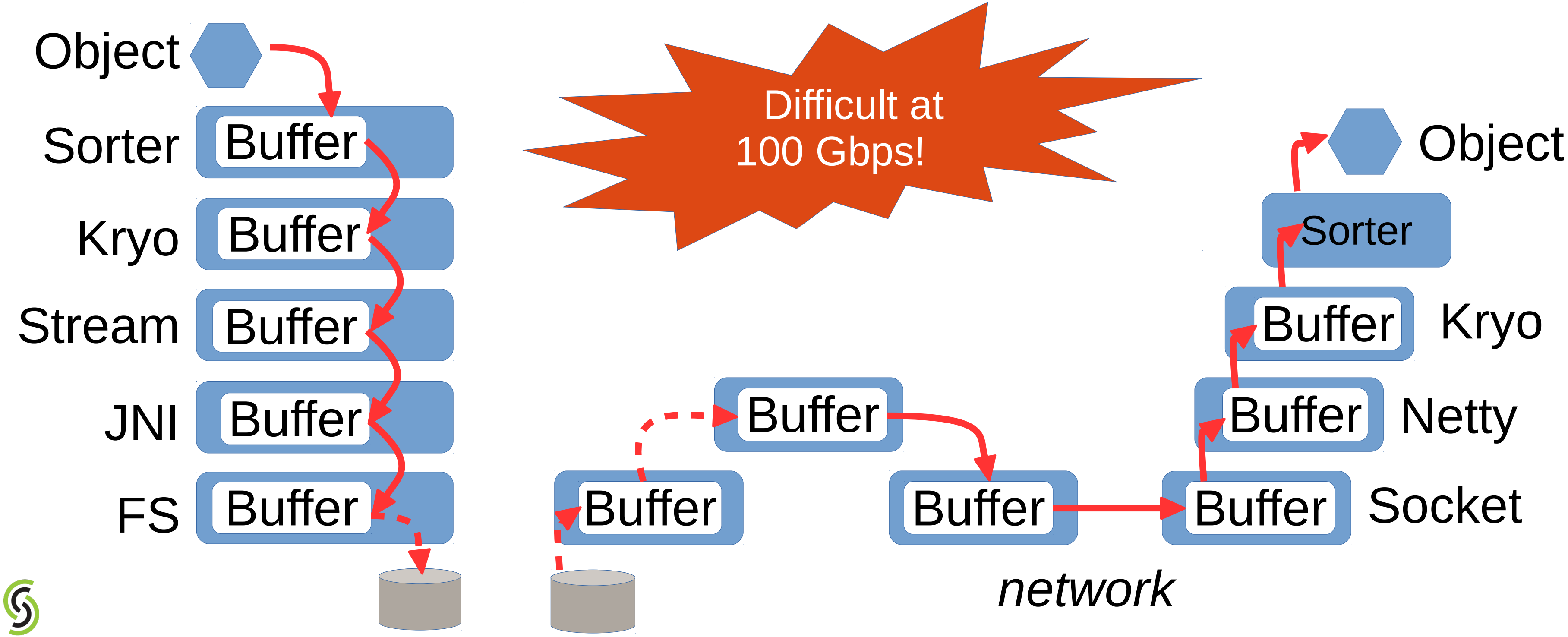
Example: Shuffle (Map)



Example: Shuffle (Map+Reduce)



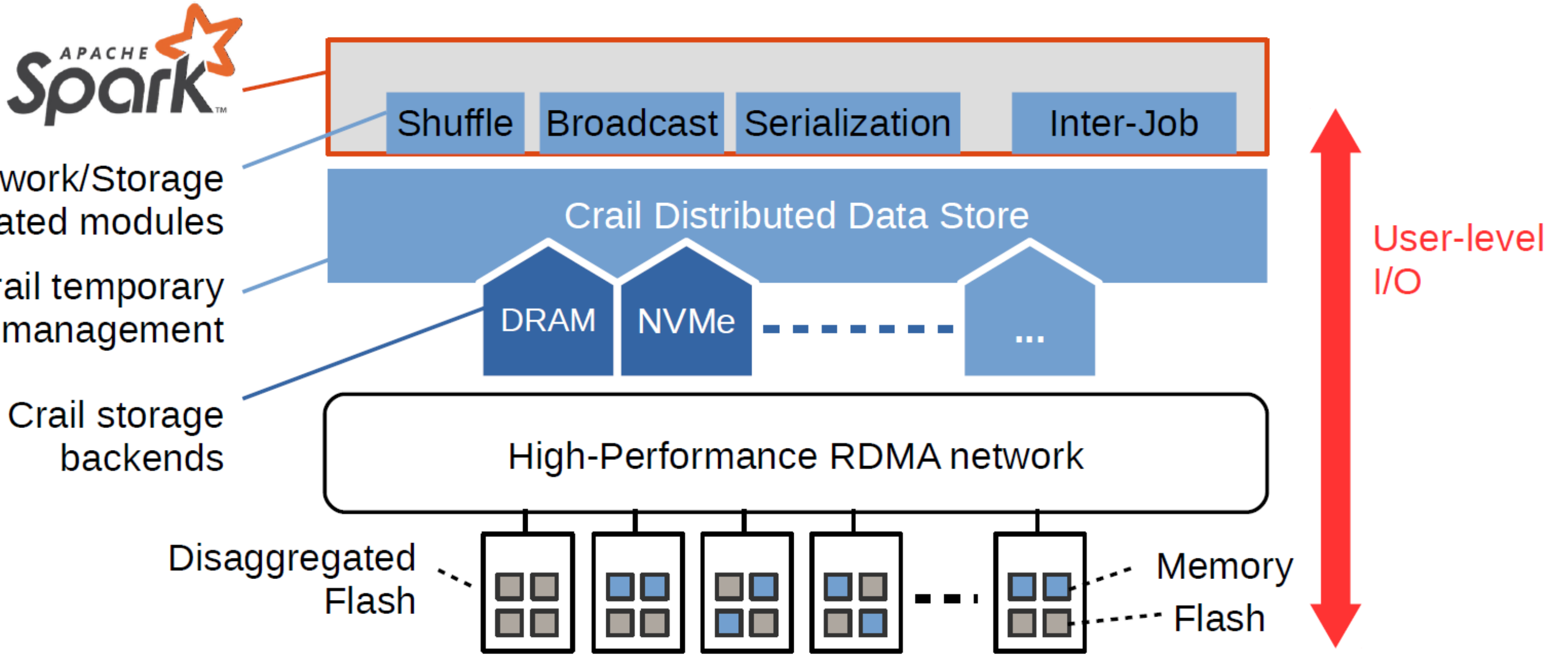
Example: Shuffle (Map+Reduce)



How can we fix this?

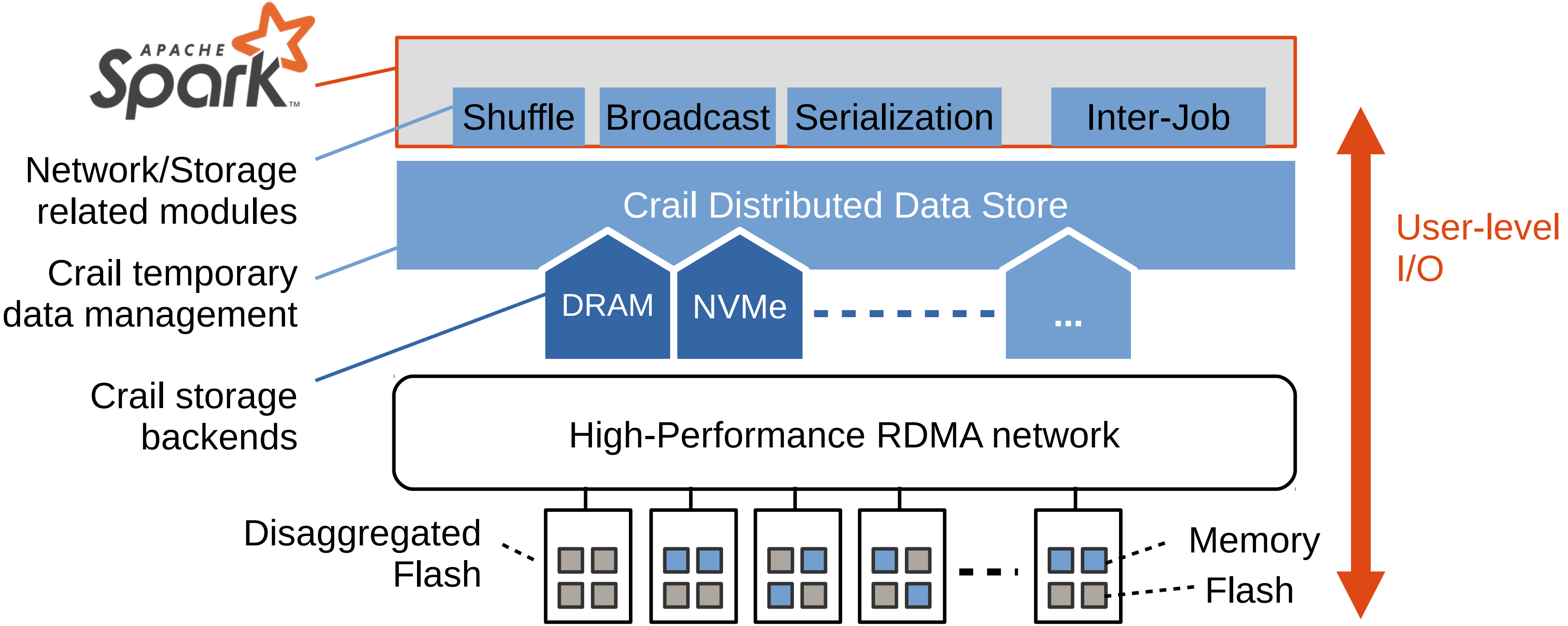
- Not just for shuffle
  - Also for broadcast, RDD transport, inter-job sharing, etc.
- Not just for RDMA and NVMe hardware
  - But for any possible future high-performance I/O hardware
- Not just for co-located compute/storage
  - Also for resource disaggregation, heterogeneous resource distribution, etc.
- Not just improve things
  - Make it perform at the hardware limit

The CRAIL Approach



Thank You.

The CRAIL Approach



Apache Spark
Shuffle
Broadcast
Serialization
Inter-Job
Network/Storage related modules
Crail Distributed Data Store
Crail temporary data management
DRAM
NVMe
...
Crail storage backends
High-Performance RDMA network
Disaggregated Flash
Memory
Flash
User-level I/O