Bringing HPC Algorithms to Big Data Platforms

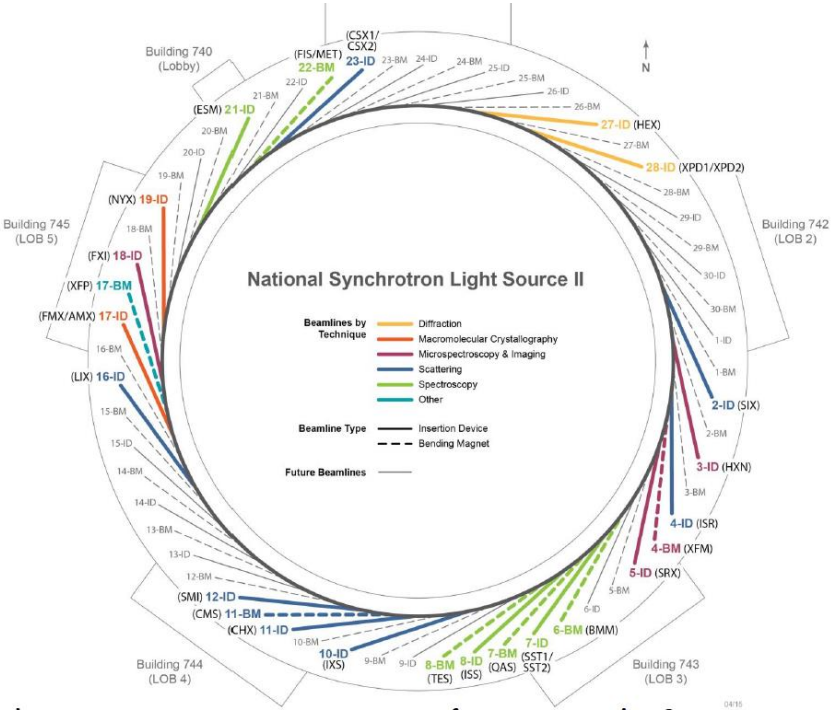
Nikolay Malitsky

Brookhaven National Laboratory

Outline

- Spark as an integrated platform for experimental facilities
- Ptychographic application
- Spark-MPI approach
- Summary

National Synchrotron Light Source II

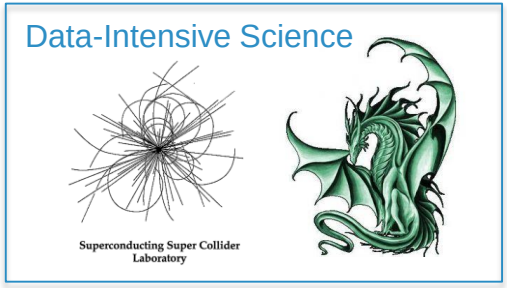


- highly optimized 3rd generation synchrotron facility
- started operations in 2014 at Brookhaven National Laboratory, New York State
- suite of six experimental programs:
  - Hard X-Ray Spectroscopy
  - Imaging & Microscopy
  - Structural Biology
  - Soft X-Ray Scattering & Spectroscopy
  - Complex Scattering
  - Diffraction & In Situ Scattering

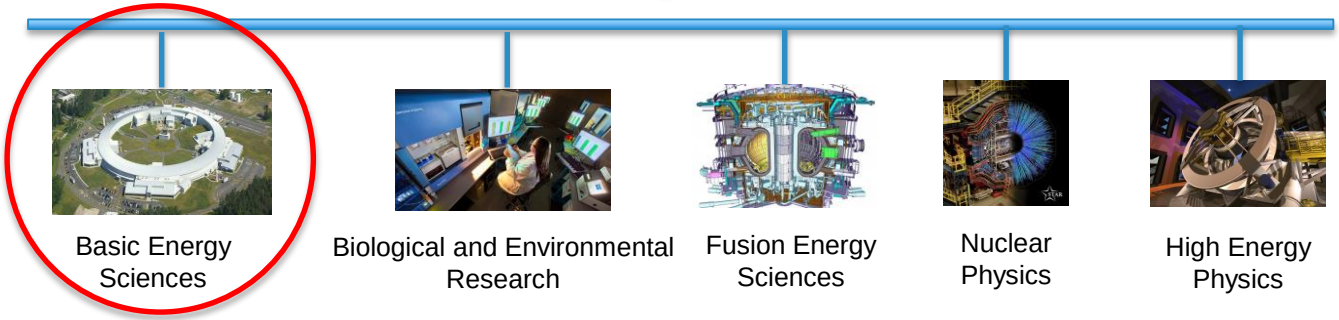
John Hill et al. NSLS-II Strategic Plan, 2015

DOE Science Drivers

Many years ago …

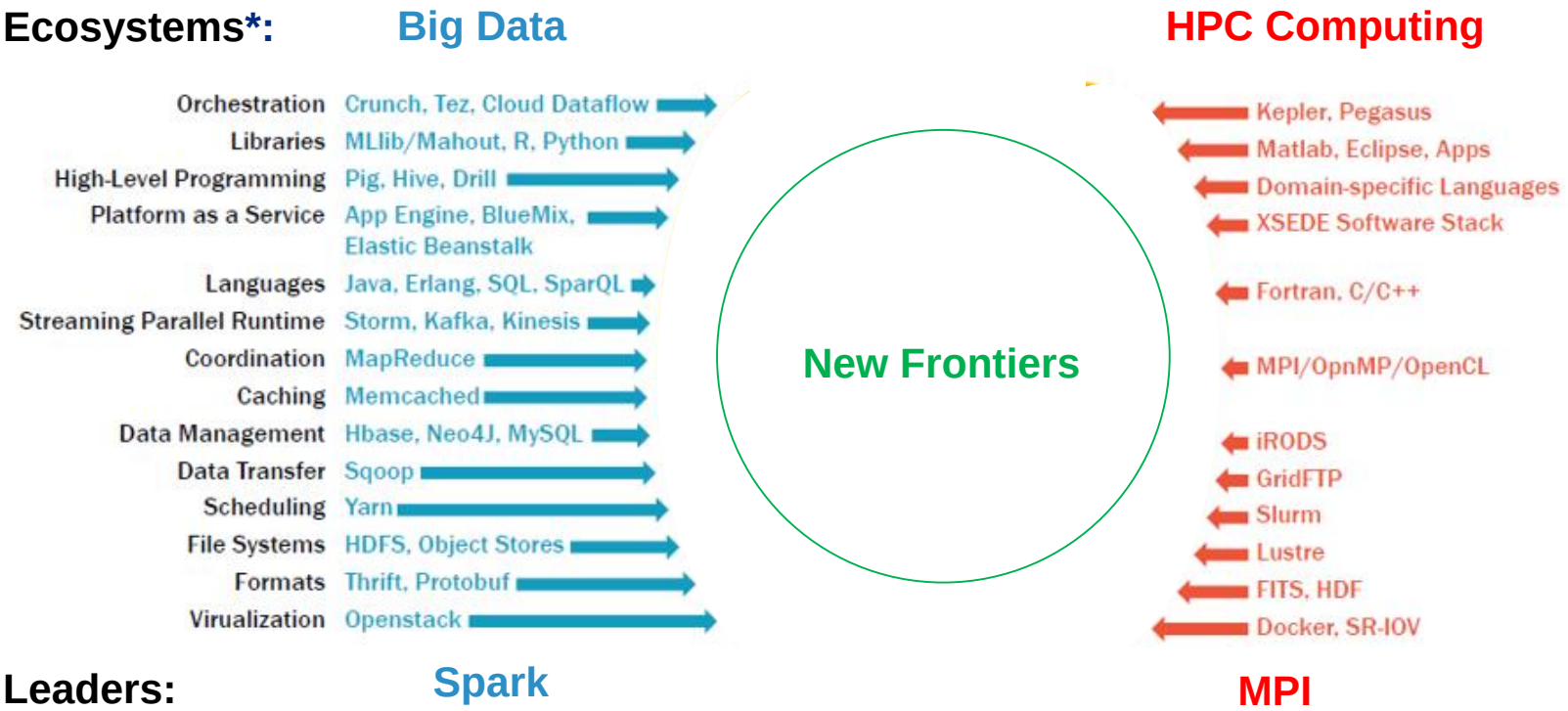


Now



NSLS-II is here

Closing a gap between Big Data and HPC computing



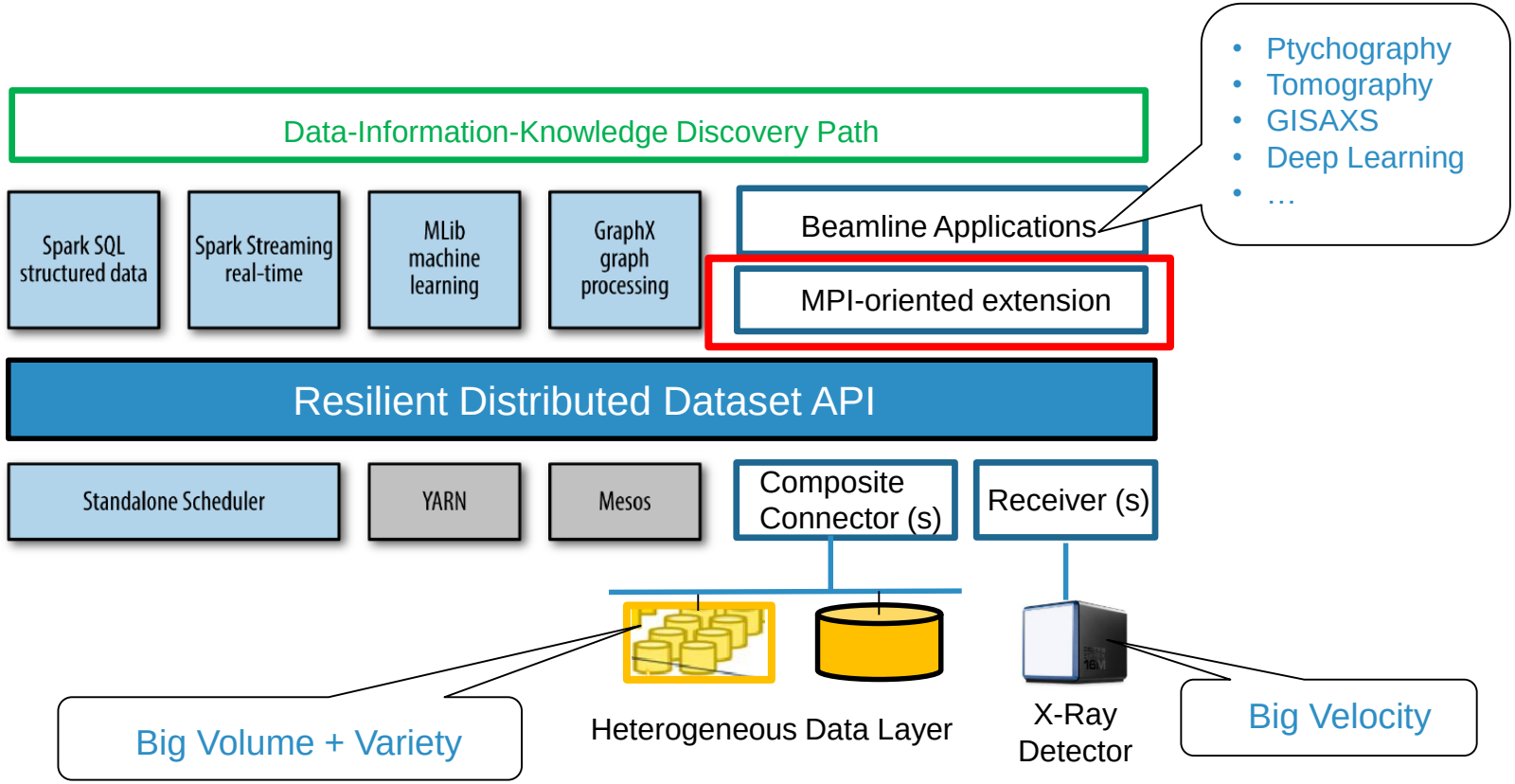
*Geoffrey Fox et al. HPC-ABDC High Performance Computing Enhanced Apache Big Data Stack, CCGrid, 2015

Three directions

- ❑ Spark + MPI-oriented extension
- ❑ MPI + Spark-oriented extension
- ❑ New model

topic of this talk

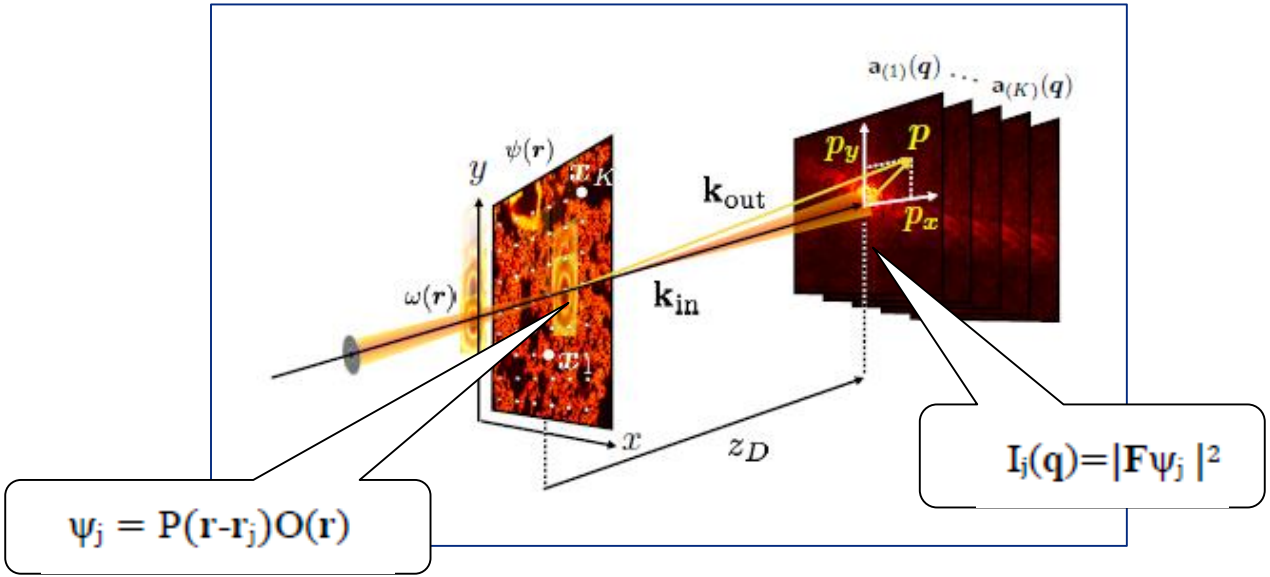
Spark an integrated platform for experimental facilities



Ptychographic Application

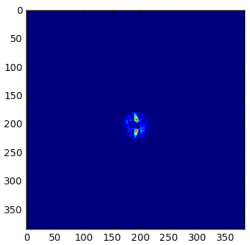
Ptychography

Ptychography is one of the essential image reconstruction techniques used in light source facilities. This method consists of measuring multiple diffraction patterns by scanning a finite illumination (also called the probe) on an extended specimen (the object).

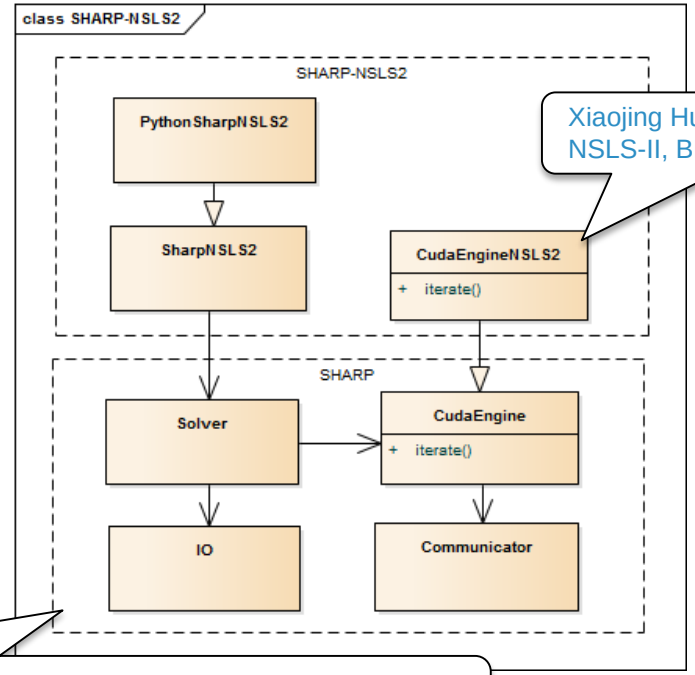


Stefano Marchesini et al. SHARP: a distributed, GPU-based ptychographic solver, J. Appl. Cryst., 49, 2016

SHARP-NSLS2 application



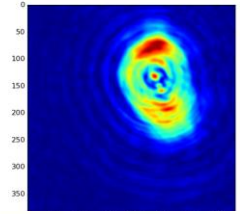
512 384x384
detector frames

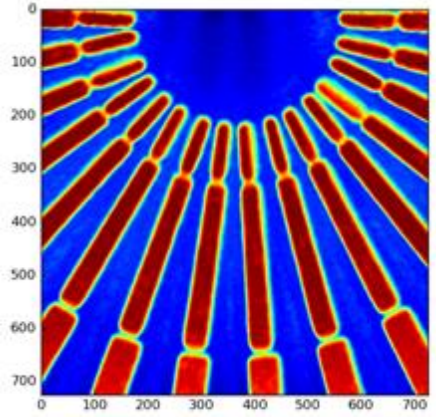


Xiaojing Huang, HXN Approach
NSLS-II, BNL

Stefano Marchesini et al. SHARP GPU-based Framework
Center for Advanced Mathematics for Energy Research, LBNL

Probe

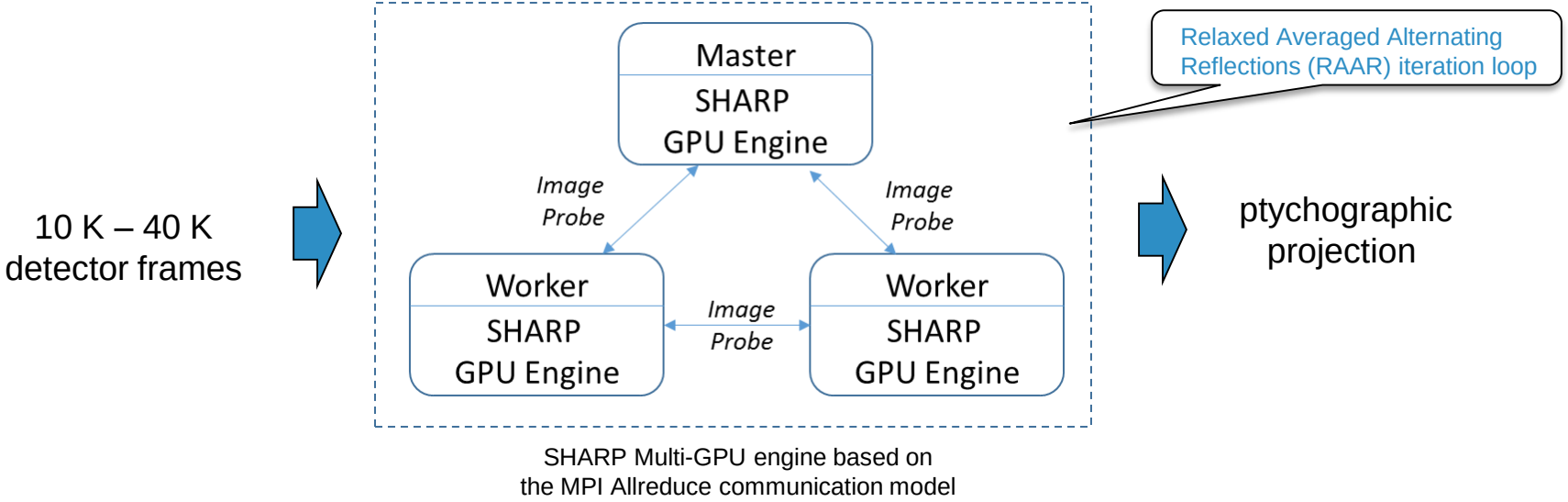




Object

Next: near-real-time ptychographic pipeline

Tomographic experiment based on 100 ptychographic projections

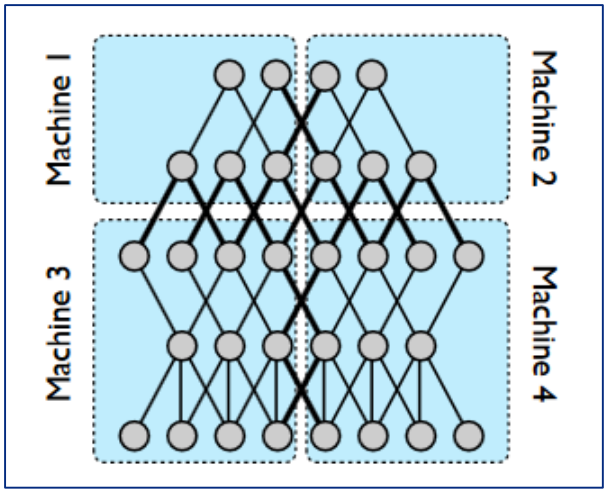


SHARP Multi-GPU engine based on
the MPI Allreduce communication model

Spark-MPI Approach

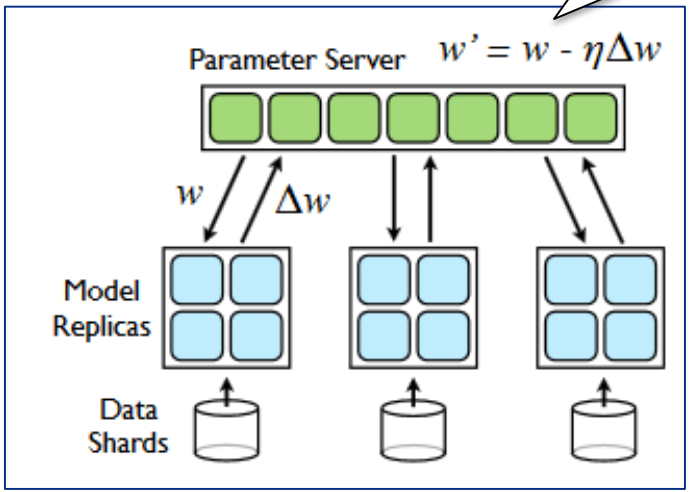
Deep Learning Parallel Approaches*

Model Parallelism



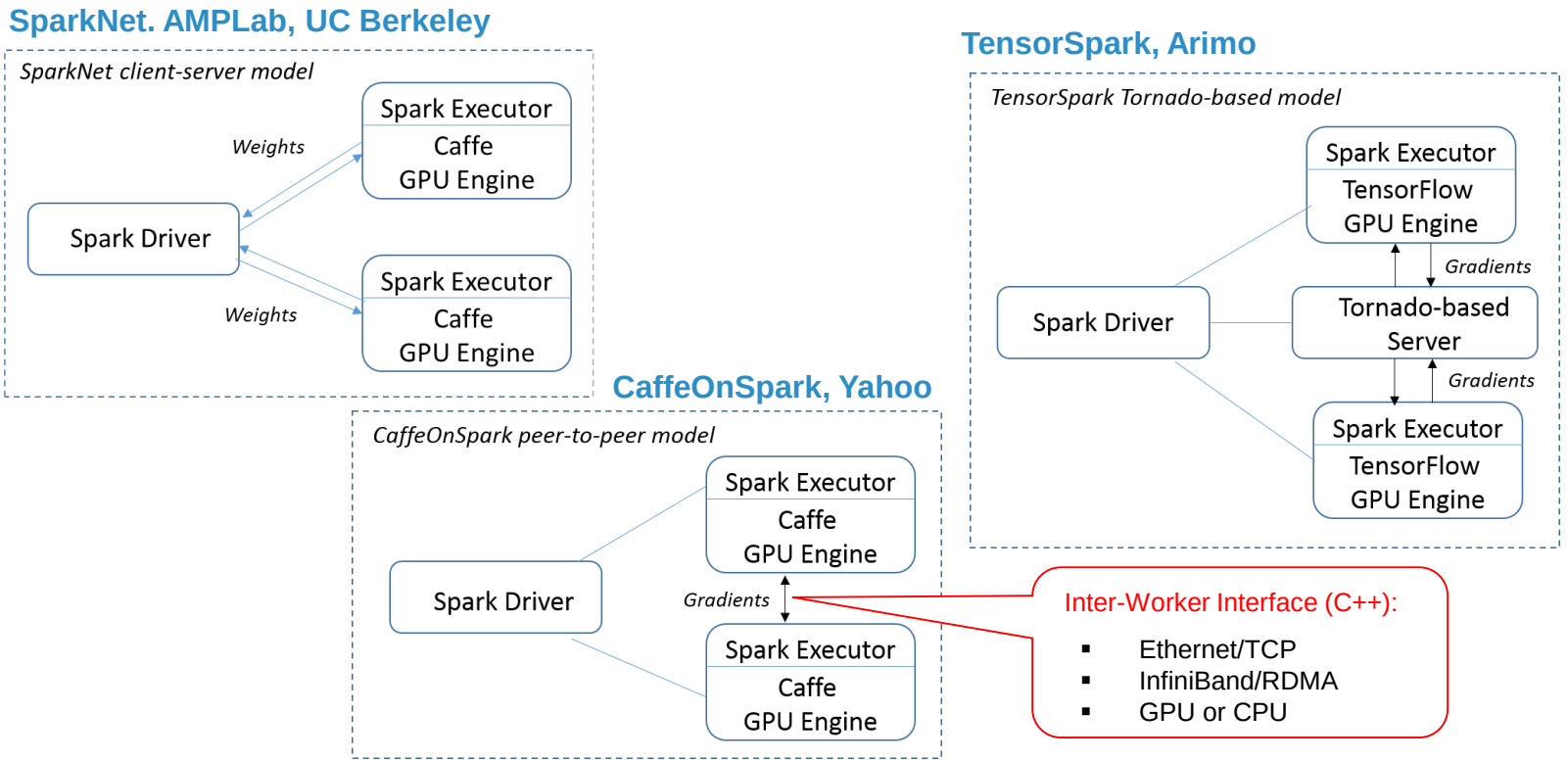
Data Parallelism

Stochastic Gradient Descent (SGD) iteration loop



*Jeffrey Dean et al. Large Scale Distributed Deep Networks, NIPS, 2012

(Some) Spark-Based Distributed Deep Learning Models*

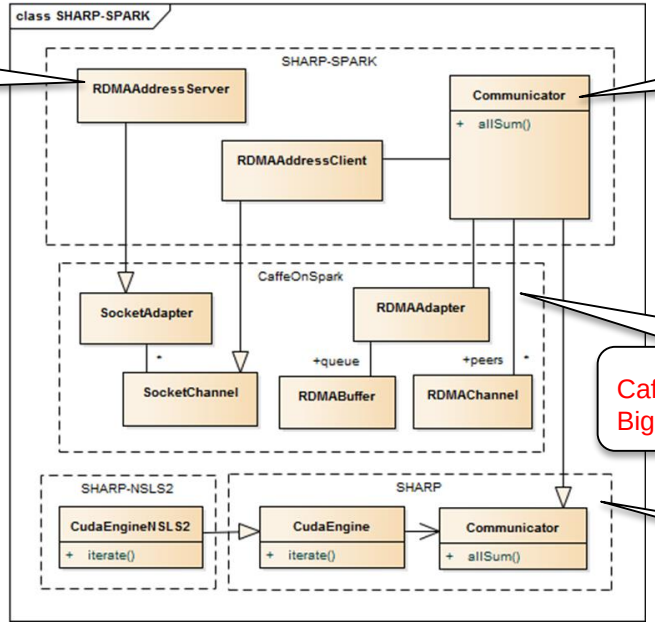


*Yu Cao and Zhe Dong. Which is Deeper – Comparison of Deep Learning Frameworks, Spark Summit, June 6-8, 2016

SHARP-SPARK Benchmark Application*

Sum of 2M arrays of floats across the Spark workers

RDMA Address Exchange Server



SHARP Inter-Worker API

CaffeOnSpark RDMA Inter-Worker Interface
Big ML Team, Yahoo

SHARP Multi-GPU Framework
CAMERA, LBNL

Benchmark results on a cluster with 8 GPU nodes equally divided among the MPI and Spark applications

Approach	Time, s
MPI Allreduce based on MVAPICH2	0.013
SHARP-SPARK based on the CaffeOnSpark library	0.016

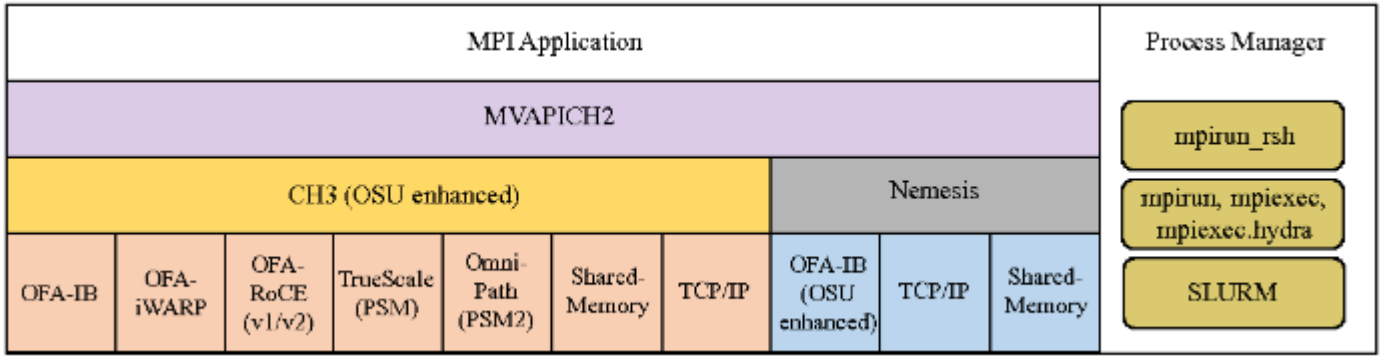
*Nikolay Malitsky, Bringing the HPC Reconstruction Algorithms to Big Data Platforms, NYSDS, August 14-17, 2016

Message Passing Interface (MPI) Framework

Major open-source implementations:

- MPICH, 1992 - present: Argonne National Laboratory
- MVAPICH, 2001 - present: Ohio State University
- OpenMPI, 2003 – present: multiple members

MVAPICH2 architecture*:



*MVAPICH Team. MVAPICH2 2.2 User Guide, 2016

From SHARP-SPARK to the MPI Framework

	SHARP-SPARK	MPI Framework
Application Programming Interface	Communicator interface	MPI-3 Standard: point-to-point, collective, etc.
Inter-Process Initialization Mechanism	RDMA address exchange server	Process Manager Interface (PMI-1 and PMI-2) with the support of several internal and external process managers
Inter-Process Communication	CaffeOnSpark RDMA library	Abstract Device Interface (ADI-3) with multiple communication adapters

Spark-MPI Conceptual Demo

https://github.com/SciDriver/spark-mpi/tree/master/examples/spark/

Create the rdd collection associated with the MPI workers

```
rdd = sc.parallelize(env, partitions)
```

Define the MPI application

```
def allreduce(kvs):
    os.environ["PMI_PORT"] = kvs["PMI_PORT"]
    os.environ["PMI_ID"] = str(kvs["PMI_ID"])

    from mpi4py import MPI

    comm = MPI.COMM_WORLD
    rank = comm.Get_rank()

    # image

    n = 2*1000000
    sendbuf = np.arange(n, dtype=np.float32)
    recvbuf = np.arange(n, dtype=np.float32)

    sendbuf[n-1] = 5.0;

    t1 = datetime.now()
    comm.Allreduce(sendbuf, recvbuf, op=MPI.SUM)
    t2 = datetime.now()

    out = {
       'rank' : rank,
       'time' : (t2-t1),
       'sum'  : recvbuf[n-1]
    }
    return out
```

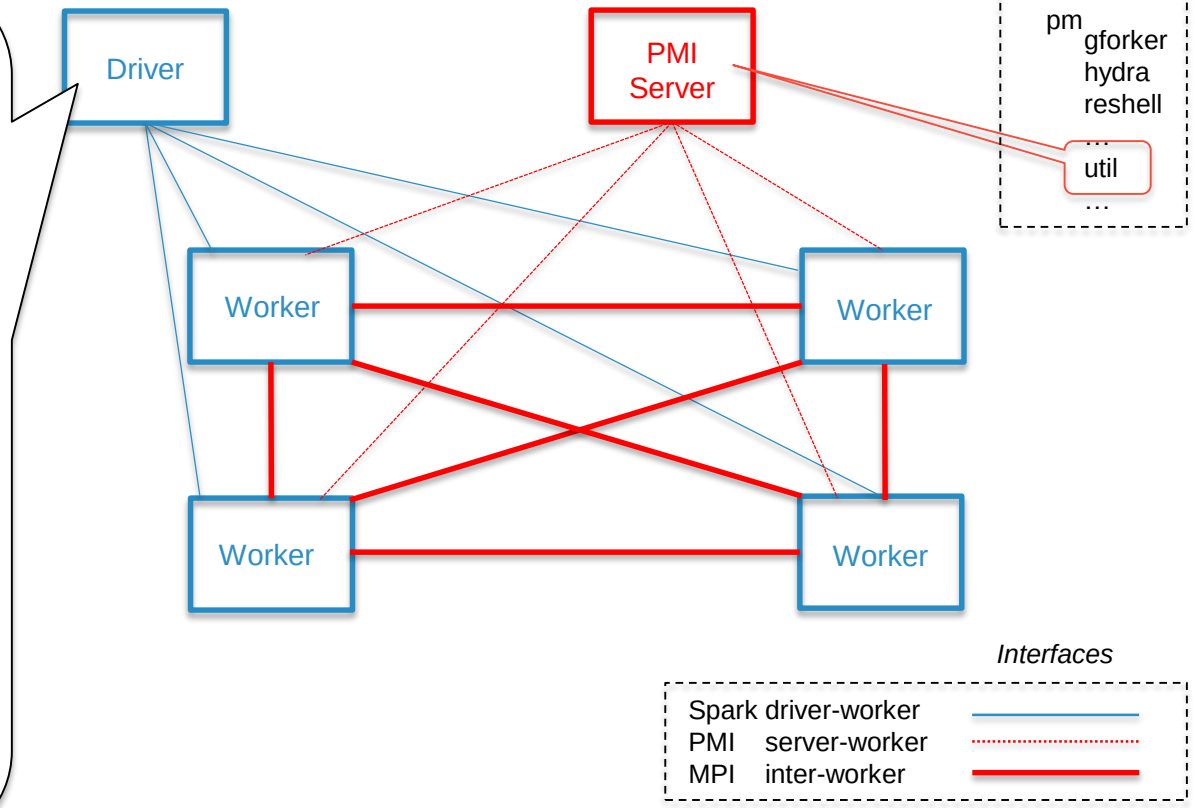
PMI Server variables

MPI interface

Run MPI application on Spark workers and collect the results

```
results = rdd.map(allreduce).collect()
for out in results:
    print ("rank: ", out['rank'], ", sum: ", out['sum'], ",
```

```
rank:  0 , sum:  20.0 , processing time:  0:00:00.014500
rank:  1 , sum:  20.0 , processing time:  0:00:00.015380
rank:  2 , sum:  20.0 , processing time:  0:00:00.014479
rank:  3 , sum:  20.0 , processing time:  0:00:00.015245
```



SPARK SUMMIT EAST 2017

Summary: Path towards the Spark-MPI Applications

- **CaffeOnSpark:** Spark + RDMA inter-worker interface + complex initialization procedure based on the Spark RDD mechanism
- **SHARP-SPARK:** Spark + CaffeOnSpark inter-worker interface + RDMA address exchange server
- **Spark-MPI:** Spark + MPI inter-worker interface + PMI Server

Kitware and BNL. An in situ, streaming, data- and compute-intensive platform for experimental data. DOE ASCR SBIR Phase I grant. Feb 21, 2017

Acknowledgement

Scientific Computing, Kitware: A. Chaudhary, P. O'Leary

CaffeOnSpark Team, Yahoo: A. Feng, J. Shi, M. Jain

SHARP Team, CAMERA, LBNL: H. Krishnan, S. Marchesini, T. Perciano, J. Sethian, D. Shapiro

Computational Science Initiative, BNL: N. D' Imperio, K. Kleese van Dam, R. D. Zhihua,

Information Technology Division, BNL: R. Perez

NSLS-II, BNL: M. Cowan, L. Flaks, A. Heroux, X. Huang, L. Li, R. Petkus, T. Smith

Funding: National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704

Thank You.

malitsky@bnl.gov