Leveraging SmartNICs in Data Management Tasks for High-Performance Computing

Craig Ulmer, Sandia National Labs, California

Matthew Curry, Sandia National Labs, New Mexico

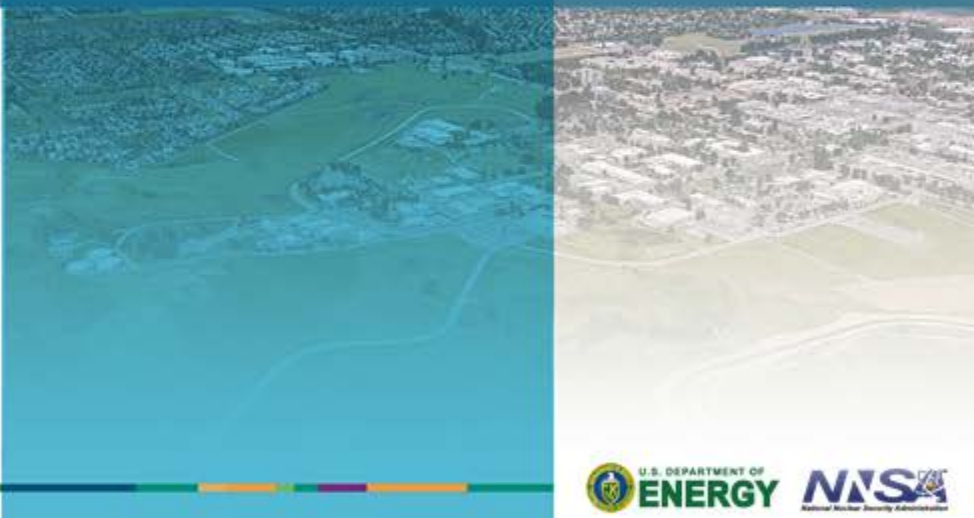
Carlos Maltzahn & Jianshen Liu, University of California Santa Cruz

U.S. DEPARTMENT OF ENERGY | Office of Science

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Field Work Proposal Number 20-023266.

UC SANTA CRUZ CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

Sandia National Laboratories



U.S. DEPARTMENT OF ENERGY | NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2021-12527 PE

Overview: Eusocial Devices in the Network Fabric

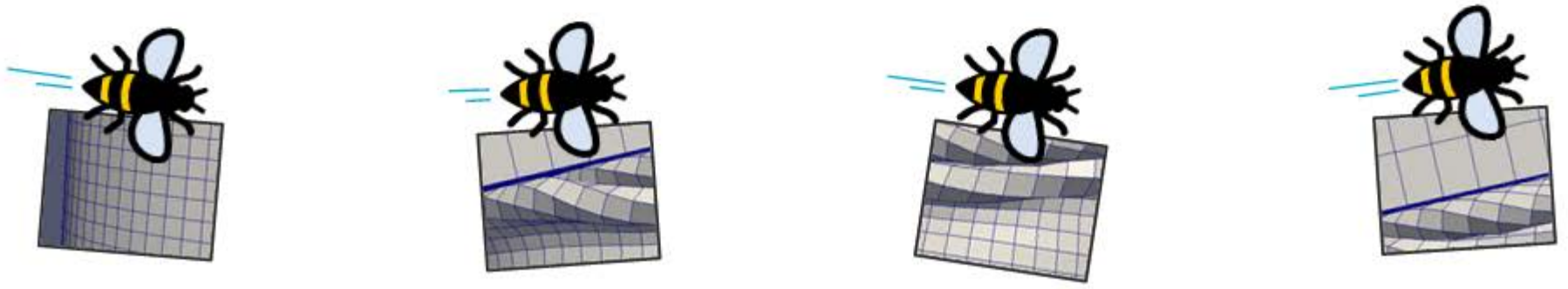
Eusocial Devices: Low-cost embedded processors allow vendors to place them everywhere
Many hands make light work

Storage: Programmable hard drives (Kinetic), Tabular extensions to Ceph (SkyhookDM)
Push filtering/transformation operations close to disks

Networks (new): Smart Network Interface Cards (SmartNICs)
Inspect and process data as it moves through network

Problem Space: High-Performance Computing simulation workflows
Generate more data than we can store, Analyze as it migrates between jobs

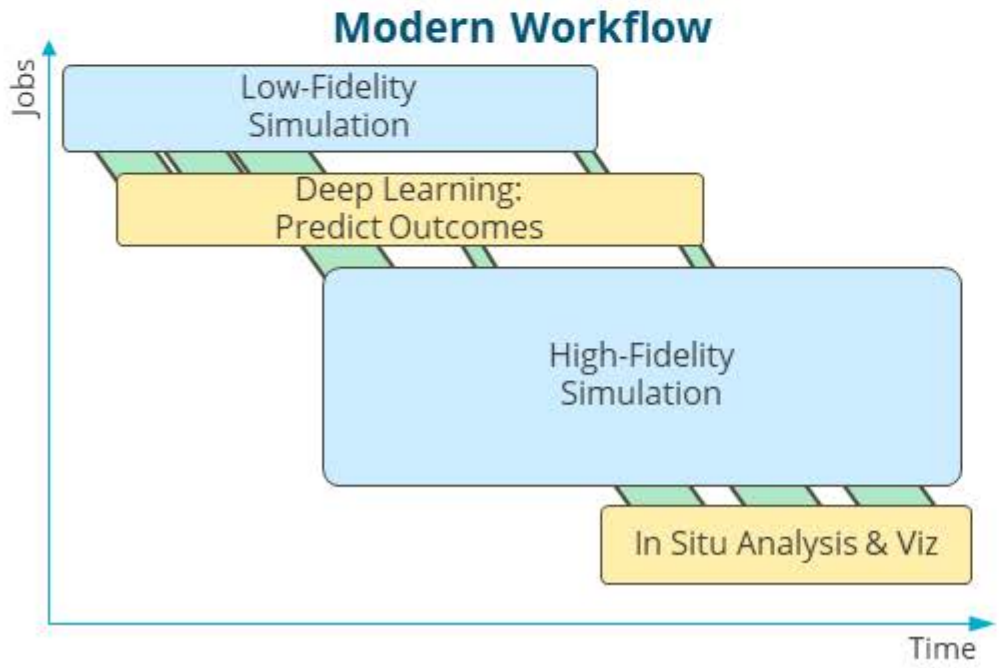
This Talk: Adapting FAODEL data management library to work with SmartNICs
Establishes a communication environment for eusocial work

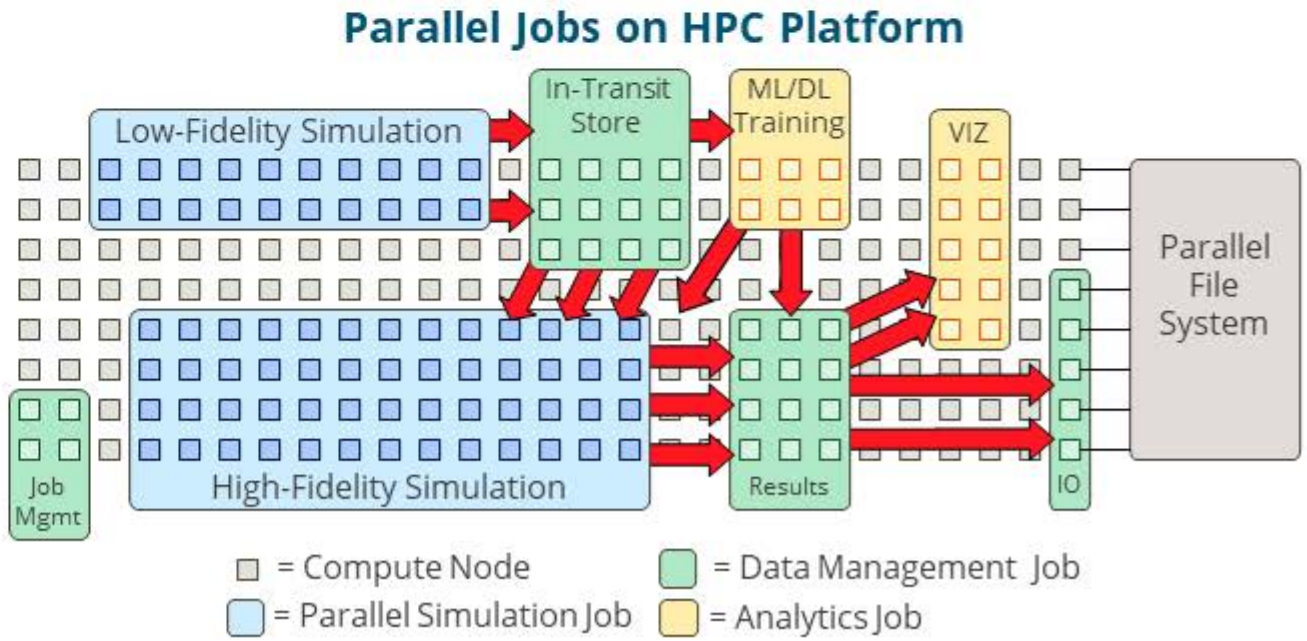


Background & Motivation

High-Performance Computing Workflows

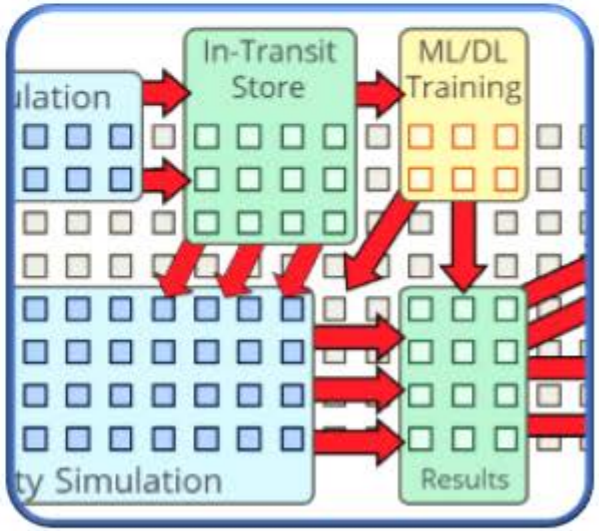
- Scientists run massively-parallel simulations to answer difficult research questions
  - Runtime datasets are too large to store (Summit ~3PB of RAM)
  - Couple different analysis tools to simulations to harvest information
- Modern workflows involve multiple parallel applications





Data Management and Storage Services

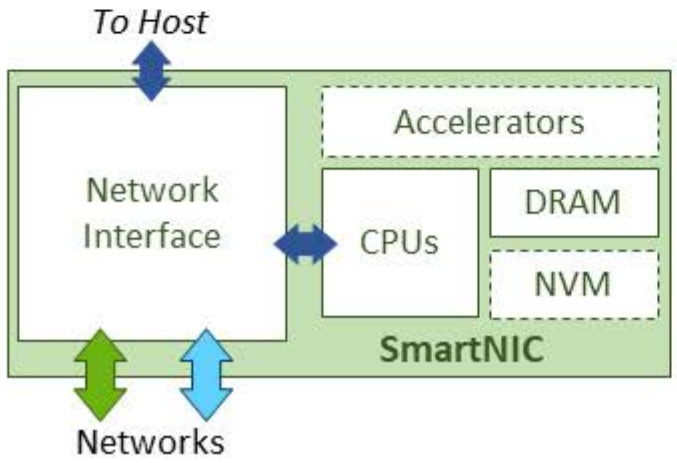
- HPC community has multiple data management libraries for routing data between jobs
  - DataSpaces, Mochi, Conduit, FAODEL
- Distributed memory services
  - Dedicate a number of nodes to serve as a pool for housing objects in memory
  - Use RDMA methods and event-driven semantics to move objects efficiently
- Problem: Services consume resources
  - Simulation Nodes: Steal cycles/memory from simulations
  - Memory Pool Nodes: Underutilize compute resources
- How can we insert cheaper memory pools?
- How can we create an environment for in-transit computations?



SmartNICs

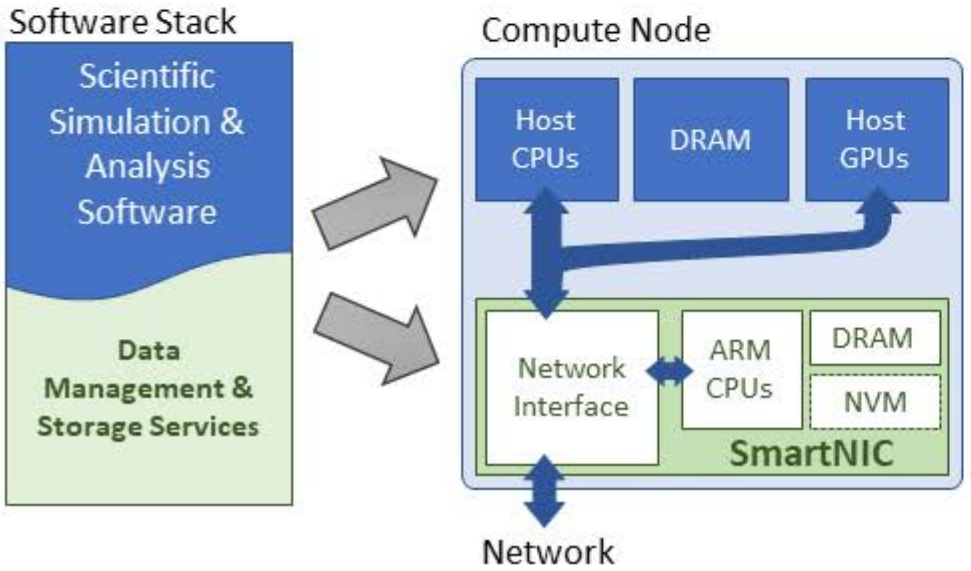
Smart Network Interface Cards (SmartNICs)

- SmartNICs are NICs that feature additional, *user-programmable* resources
  - Embedded ARM processors, DRAM, NVM, accelerators
  - Cloud Computing (AWS and Azure) and network security driving market
  - Vendors: NVIDIA, Fungible, Intel, Xilinx
- NVIDIA BlueField Timeline
  - BlueField-2 (2021): 8 ARM cores, 16GB RAM, 16GB Storage, 100Gb/s Ethernet/InfiniBand
  - BlueField-3 (2022): "10x improvement"
  - BlueField-4 (2024): "100x improvement"

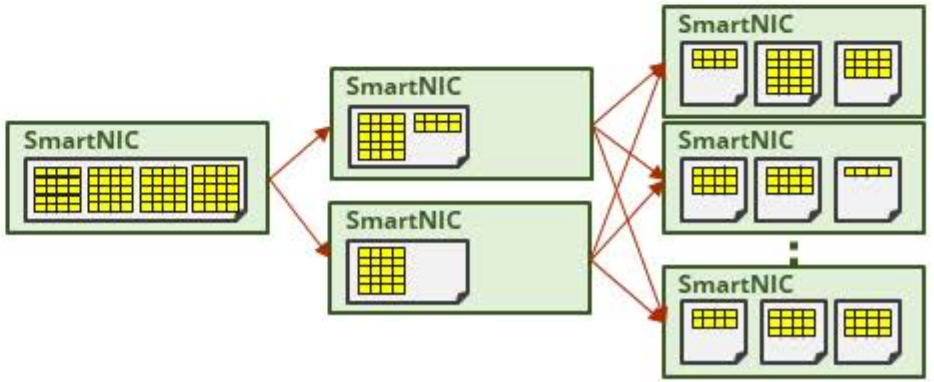


Leverage SmartNICs to Offload Services

- Split software stack to allow data management services to be offloaded to Network's Edge
- Long-Term Project Objectives
  - Reduce number of compute nodes required to store objects
  - Improve the rate at which simulations can publish objects
  - Provide environment for executing computations on in-transit data waves
  - Construct indexing services for users to perform queries on distributed tabular data



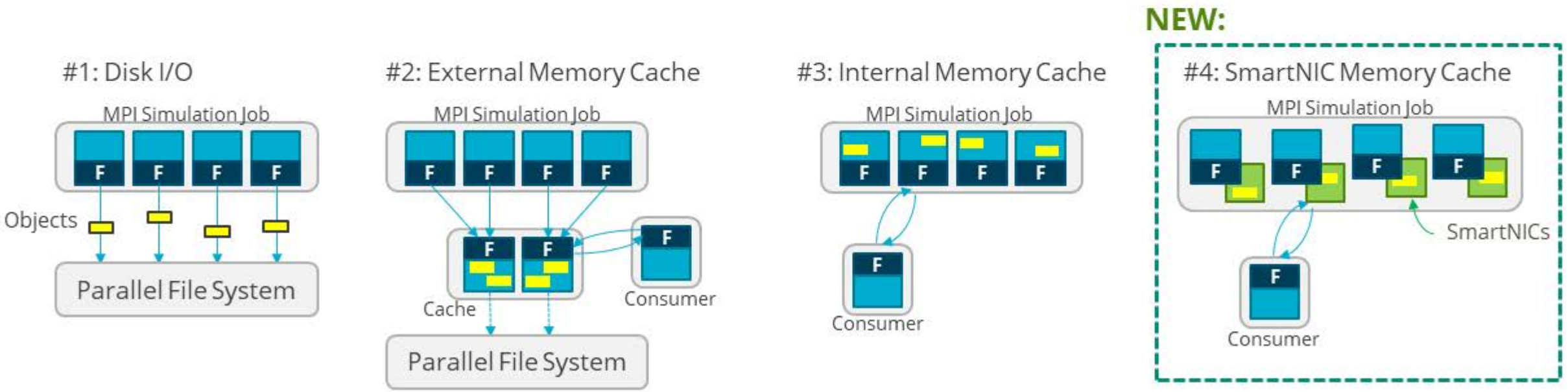
Tabular Workflows Example: Log-Structured Merge Trees



HPC Data Management Services

FAODEL: Flexible, Asynchronous, Object Data-Exchange Libraries

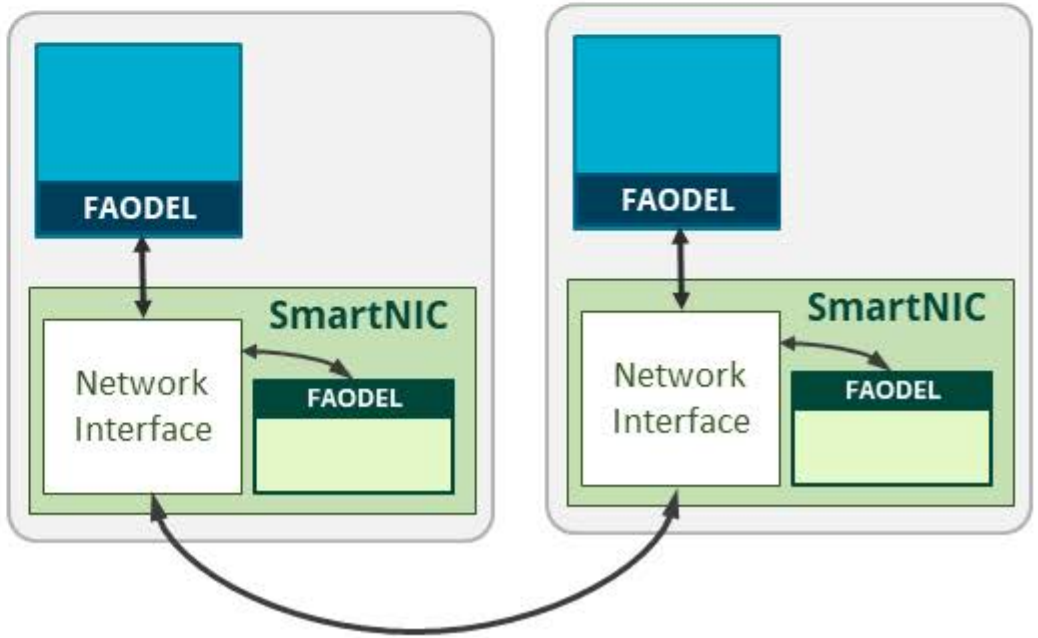
- Sandia leverages FAODEL library for routing objects between workflow jobs
  - FAODEL uses safe RDMA primitives to move objects between applications and pools
  - Objects can be stored on disk, in external memory, or in original producer's internal memory
- SmartNIC Opportunity: Host objects on node, but close to network fabric
  - Simplifies host management obligations, improves reliability of workflow



https://github.com/faodel

Adapting FAODEL to run on both Hosts and BlueField SmartNICs

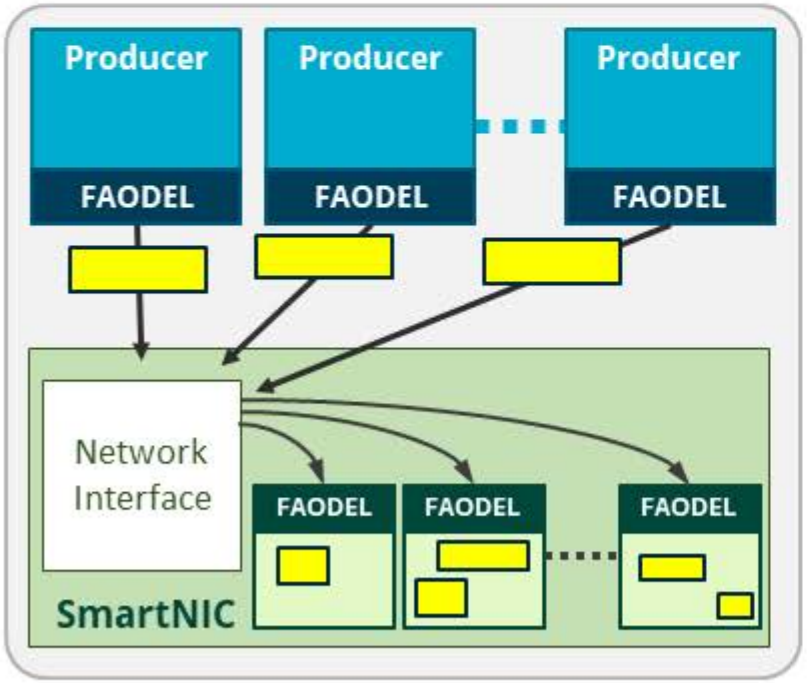
- Did not require conceptual changes to FAODEL
  - BlueField ARMs appear as just another compute node in network fabric
  - Pool abstraction allows any endpoint to participate in one or more pool
  - Users add BlueField nodes to the runtime pool configuration to use them for hosting data
- Transitioning to ARM challenges
  - Resolved assembly issues (tcmalloc)
  - Previously ported FAODEL to ARM platforms
- First time mixing x86_64 and ARM endpoints

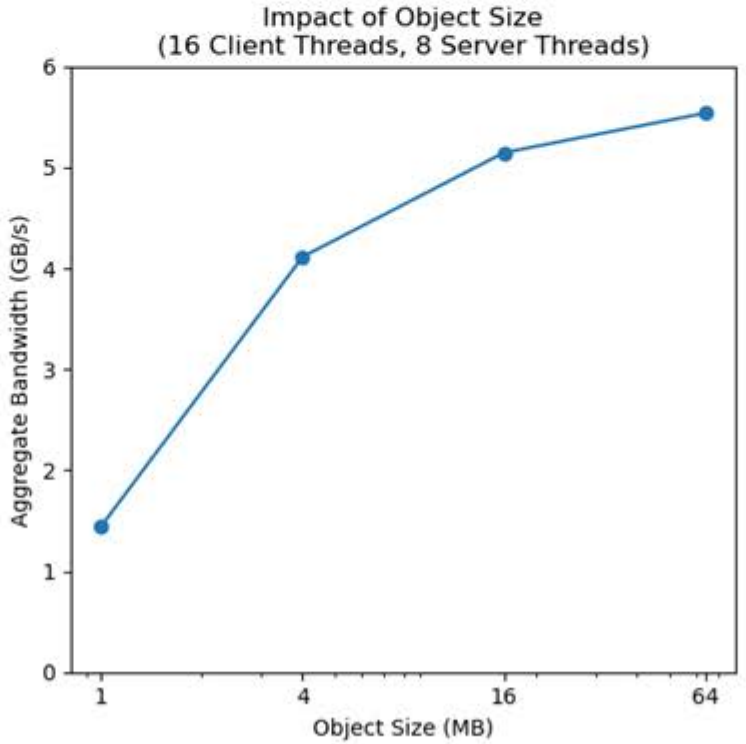


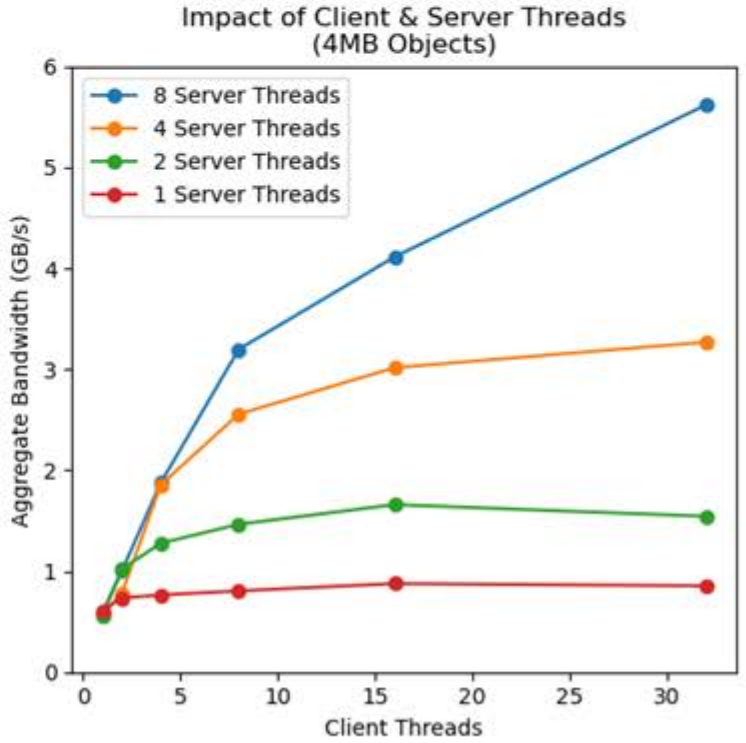
Performance Experiments

FAODEL Injection Tests

- How fast can the host push objects to FAODEL pool on local SmartNIC?
  - Injection test uses multiple threads to send objects to card at the same time
  - Report aggregate bandwidth from *application perspective* (allocate, copy, transfer, acknowledge)
  - Utilized BlueField-1 for this work







Takeaway: Increasing object size and threads improves performance. Overhead because not leveraging locality.

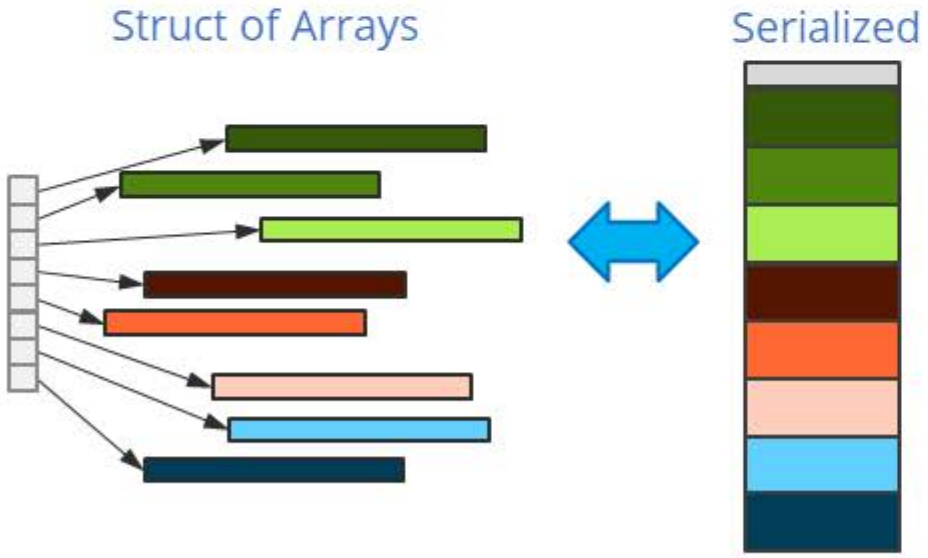
FAODEL Stress Experiments

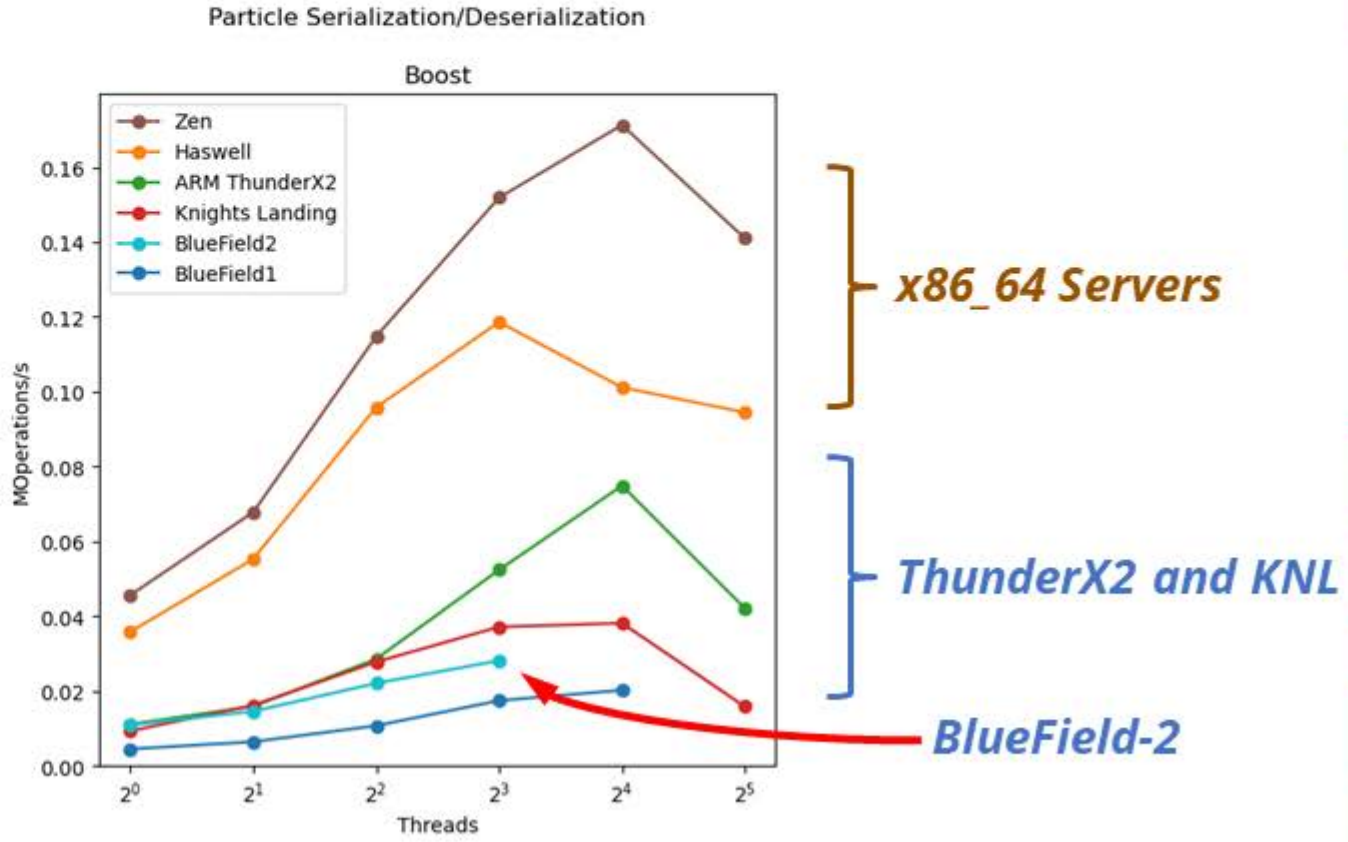
- How are data management tasks impacted by embedded processors?
  - Stress-ng benchmark inspired us to create **faodel-stress** tool
  - Generating/sorting keys, serializing data, allocating network memory, hash maps, …
  - Compared BlueField to a variety of servers used today in HPC
- Two Examples: Particle serialization and Local Key/Blob store

Processor	Year	Architecture	Cores	Frequency	Memory
Zen	2017	AMD EPYC 7601	2x32	2.2 GHz	1024 GB
Haswell	2016	Intel E5-2698 v3	2x16	2.3 GHz	128 GB
Knights Landing (KNL)	2016	Intel Phi 7250	1x68	1.4 GHz	16+96 GB
ARM ThunderX2	2018	ARM Thunder X2	2x28	2.5 GHz	128 GB
BlueField-2	2021	ARMv8 A72	1x8	2.5 GHz	16 GB
BlueField-1	2018	ARMv8 A72	1x16	800 MHz	16 GB

FAODEL Stress Experiments: Particle Serialization

- Serialization: Convert user's data structures to contiguous buffers
  - Being able to unpack, analyze, and repack data is a critical function for eusocial work
  - Test: Particle bundle data (8 x 1K)

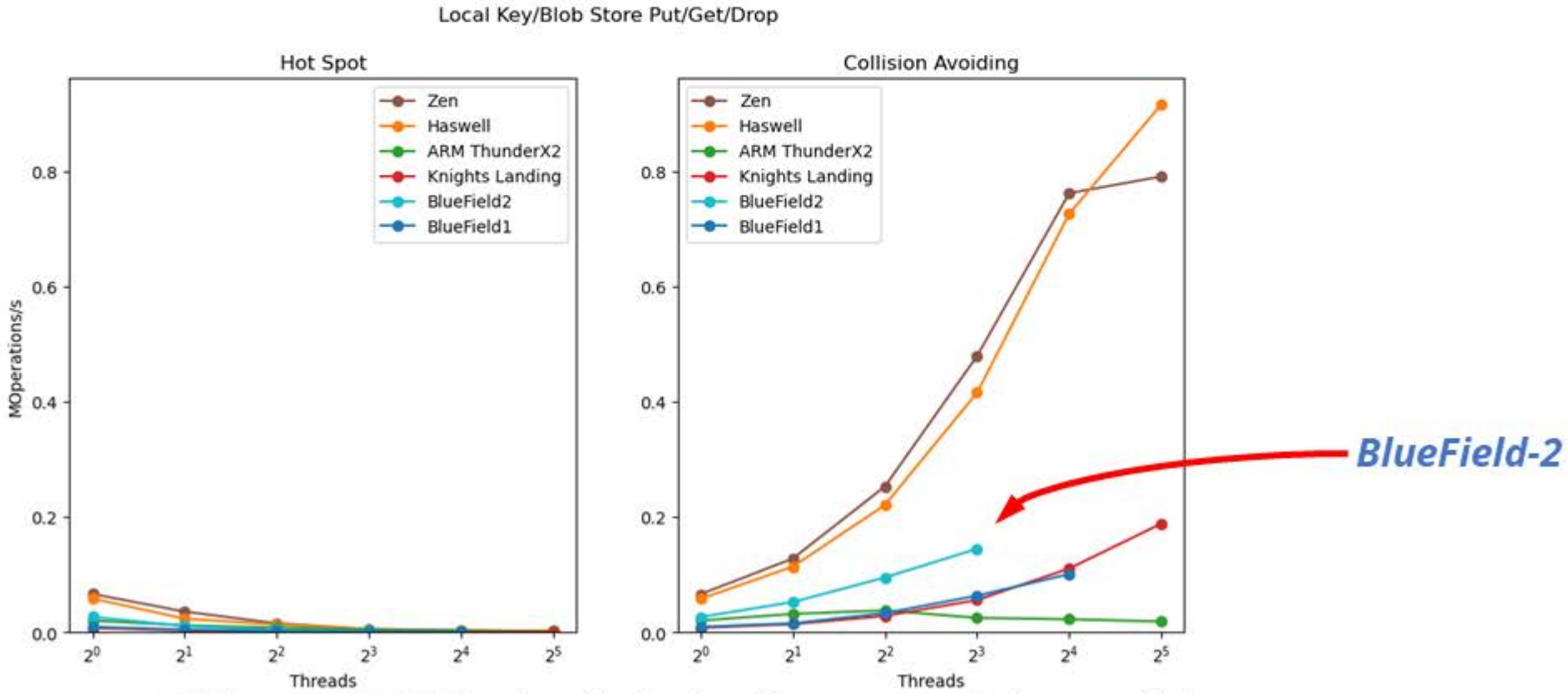




Takeaway: Two distinct performance classes, BlueField on lower end of performance.

FAODEL Stress Experiments: In-memory Key/Blob Store

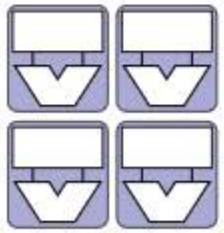
- Data structure for organizing objects and scaffolding for event-driven operations
  - Perform put/get/drop operations in rapid succession
  - Use key names that either create or avoid contention



Takeaway: BF2 actually faster than some data-parallel processors.

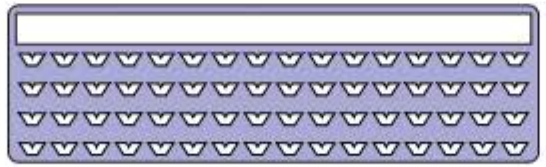
Challenges with Data-Parallel Processors

- HPC processors are picked to optimize compute performance of simulations
  - Goal: place as many floating-point units in a chip as possible
  - KNL And ThunderX2 use many cores with *lower individual-core* performance
  - GPUs are the extreme case for this scenario
- Many data management tasks do not have the right data parallelism to exploit these processors
- Useful to have different processors in the platform
  - Pick a compute-optimized processor for the simulation
  - Pick an embedded processor for coordinating asynchronous network operations

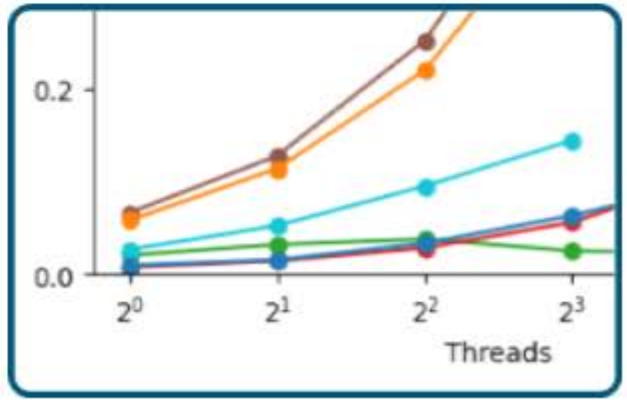


Multicore

Many Cores
(but slower)



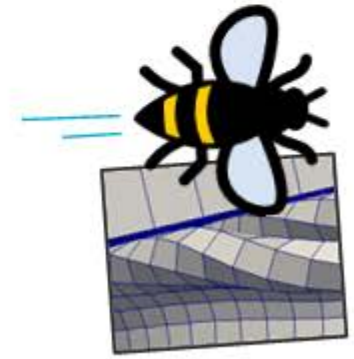
GPUs



Summary & Looking Forward

Summary & Looking Forward

- SmartNICs are appealing for HPC data management services
  - Cost-effective way to offload in-memory object hosting
  - Demonstrated we can use them with existing communication libraries
- Current hardware has limited computational performance
  - Affects throughput, limits operations that can be performed on in-transit data
  - Data parallel processors in HPC have similar problems!
  - Expect next year's BlueField-3 hardware to address some of our performance issues
- Upcoming work
  - Optimizations to improve host→card injection times
  - FAODEL's new remote computing operations API for user-defined functions
  - Leverage Apache Arrow to store and process tabular data in SmartNICs



https://github.com/faodel