Offloading Data Management Services to SmartNICS

Craig Ulmer, Sandia National Labs, California

Jianshen Liu, Aldrin Montana, & Carlos Maltzahn, University of California Santa Cruz

Matthew L. Curry, Scott Levy, & Whit Schonbein, Sandia National Labs, New Mexico

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

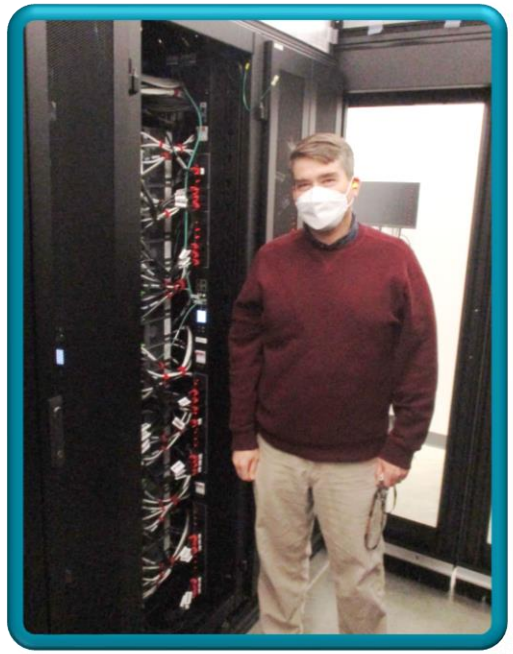
U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

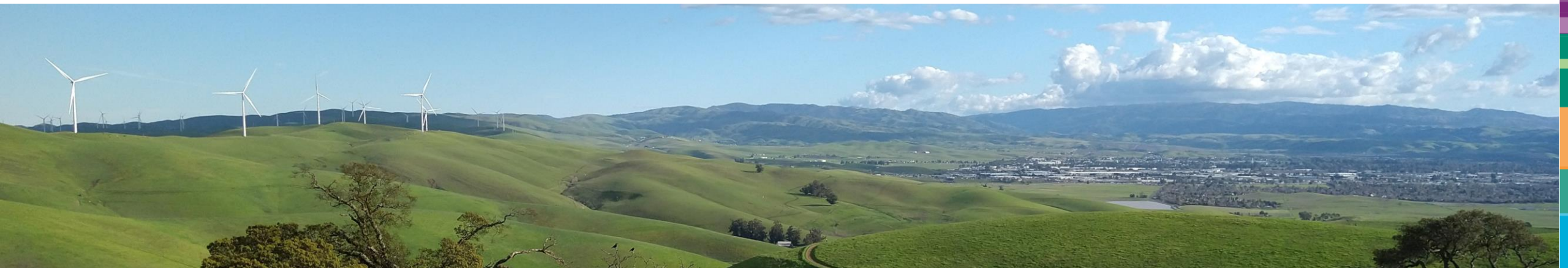
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2023-03752C

About Craig Ulmer

- Computer Engineer from Georgia Tech
  - Ph.D. for parallel computing communication software
  - Internships at IBM, Eastman Kodak, and NASA JPL
- Joined Sandia National Labs in Livermore, California (2003)
  - FPGA Accelerators: Network packet filters
  - Scientific Computing: Connecting simulations & analytics
  - Nonproliferation: Sifting through large geospatial datasets
  - ➢ SmartNICs: Leveraging SmartNICs for data management services





Outline for Today

- What is the Department of Energy? What does Sandia do?
- How do researchers use High-Performance Computing to solve problems?
- How can SmartNICs improve workflows on HPC Platforms?
  - SmartNICs
  - Creating an environment for hosting services on SmartNICs
  - Particle sifting example
- Jobs at the labs
  - https://www.energy.gov/jobs-national-labs
  - https://www.sandia.gov/careers/

What is the Department of Energy? What does Sandia do?

The Department of Energy

Manhattan Project 1942

Atomic Energy Commission 1946

Department of Energy 1977

The mission of the Energy Department is to ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.

Energy

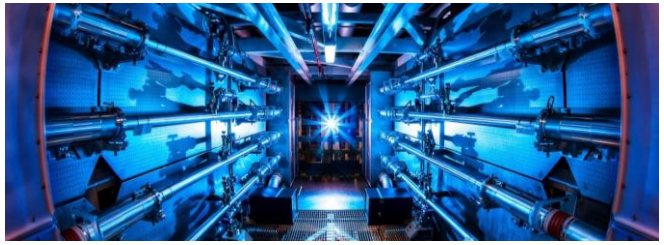
Catalyze the timely, material, and efficient transformation of the nation's energy system and secure U.S. leadership in energy technologies.

Science & Innovation

Maintain a vibrant U.S. effort in science and engineering as a cornerstone of our economic prosperity with clear leadership in strategic areas.

Nuclear Safety & Security

Enhance nuclear security through defense, nonproliferation, and environmental efforts.



The DOE has 17 National Laboratories

Office of Science Laboratories

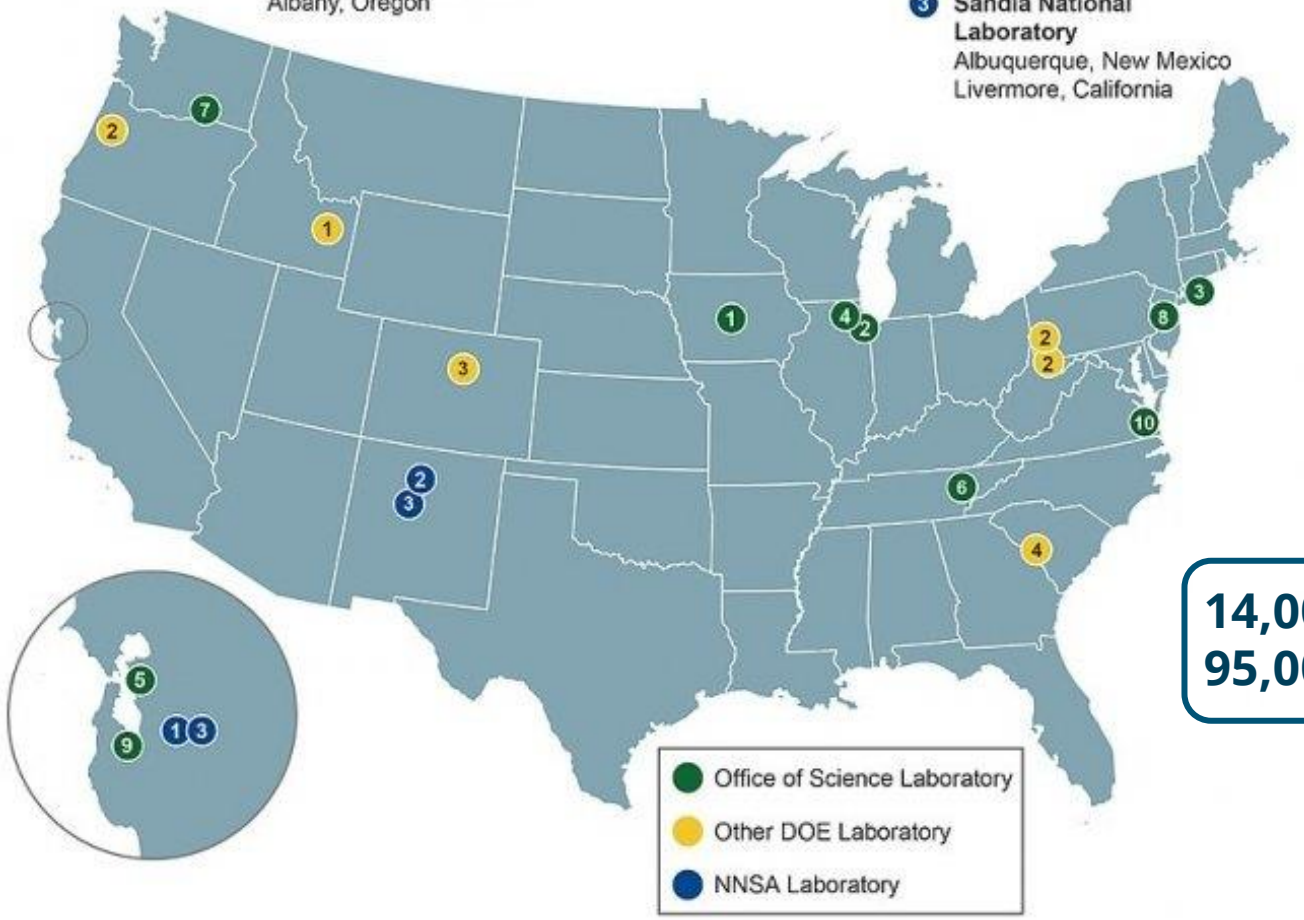
1. **Ames Laboratory**
   Ames, Iowa
2. **Argonne National Laboratory**
   Argonne, Illinois
3. **Brookhaven National Laboratory**
   Upton, New York
4. **Fermi National Accelerator Laboratory**
   Batavia, Illinois
5. **Lawrence Berkeley National Laboratory**
   Berkeley, California
6. **Oak Ridge National Laboratory**
   Oak Ridge, Tennessee
7. **Pacific Northwest National Laboratory**
   Richland, Washington
8. **Princeton Plasma Physics Laboratory**
   Princeton, New Jersey
9. **SLAC National Accelerator Laboratory**
   Menlo Park, California
10. **Thomas Jefferson National Accelerator Facility**
    Newport News, Virginia

Other DOE Laboratories

1. **Idaho National Laboratory**
   Idaho Falls, Idaho
2. **National Energy Technology Laboratory**
   Morgantown, West Virginia
   Pittsburgh, Pennsylvania
   Albany, Oregon
3. **National Renewable Energy Laboratory**
   Golden, Colorado
4. **Savannah River National Laboratory**
   Aiken, South Carolina

NNSA Laboratories

1. **Lawrence Livermore National Laboratory**
   Livermore, California
2. **Los Alamos National Laboratory**
   Los Alamos, New Mexico
3. **Sandia National Laboratory**
   Albuquerque, New Mexico
   Livermore, California



14,000 Federal Employees
95,000 Contractors

Sandia: Fulfilling Our National Security Mission

Global Security

Nuclear Deterrence

National Security Programs

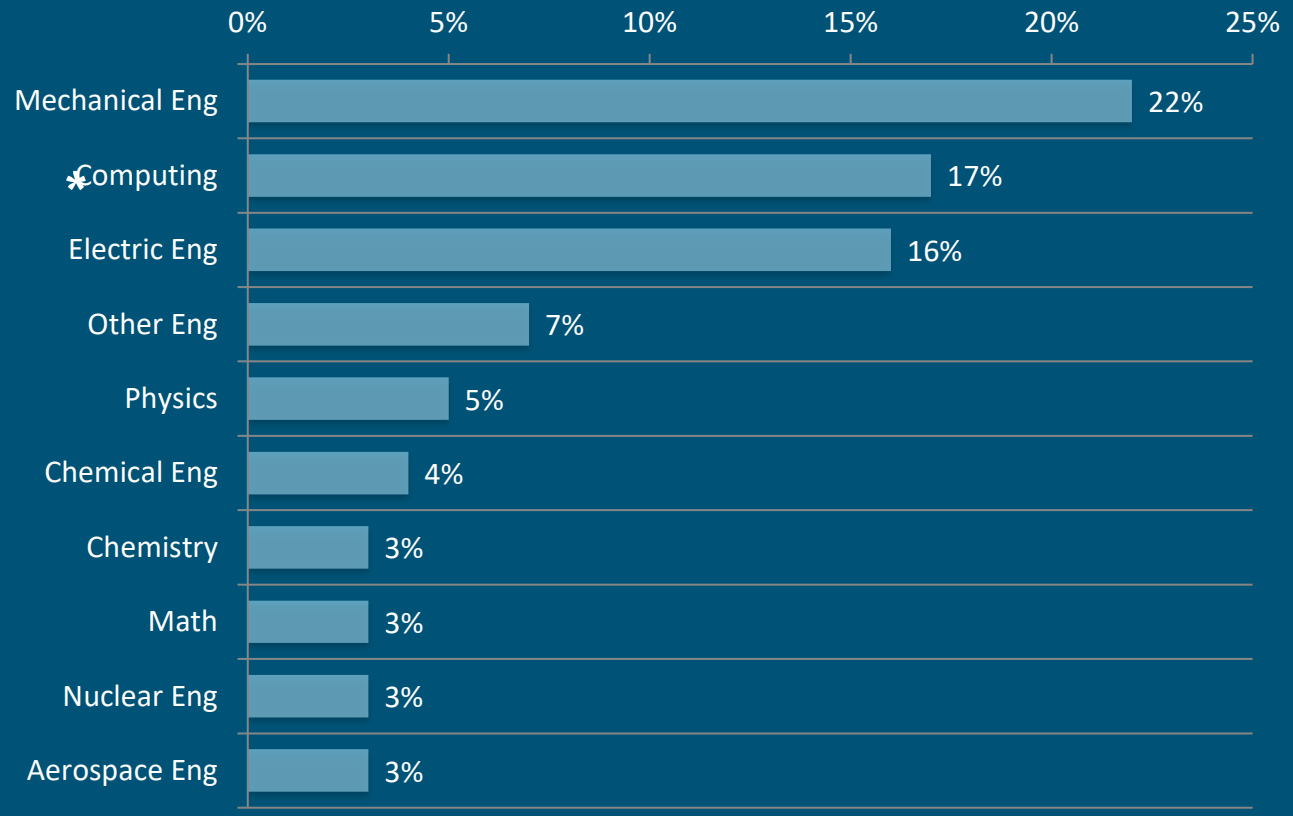
Energy & Homeland Security

Advanced Science & Technology

Some of the critical national security issues that we address lie in the cyber area.

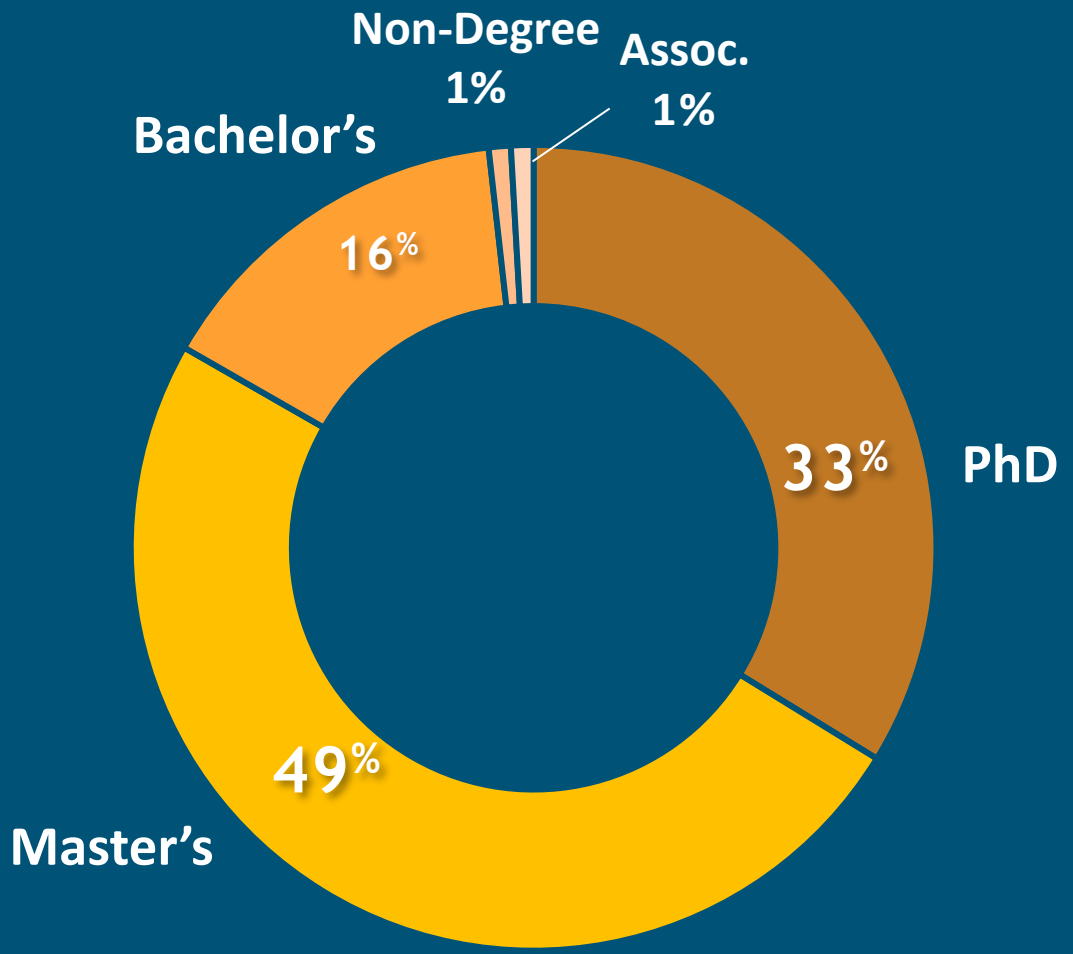
Sandia: R&D by Discipline & Degree

Sandia has approximately 16,000 Employees
14,000 NM, 2,000 CA



Top 10 job descriptions shown, Regular exempt non-management employees only

(*) Computing includes disciplines such as High Performance Computing, Cybersecurity, Machine Learning, Autonomous Sensing and Perception



Data as of July 2023

How do researchers use High-Performance Computing to solve problems?

Stakeholders Have Difficult Questions

How far apart should windmills be placed?

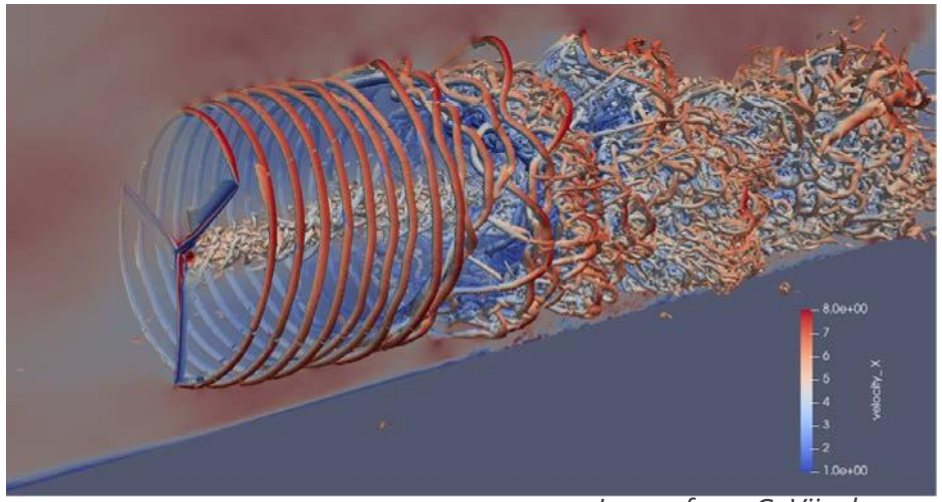
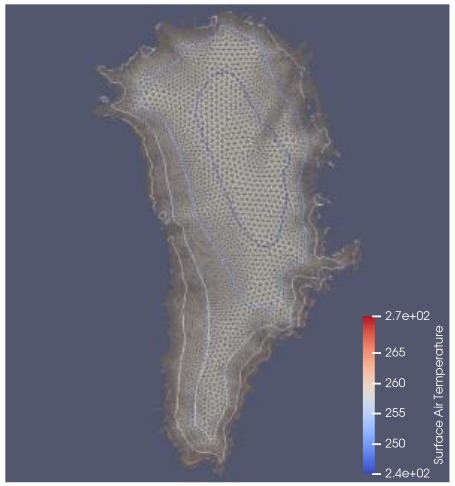
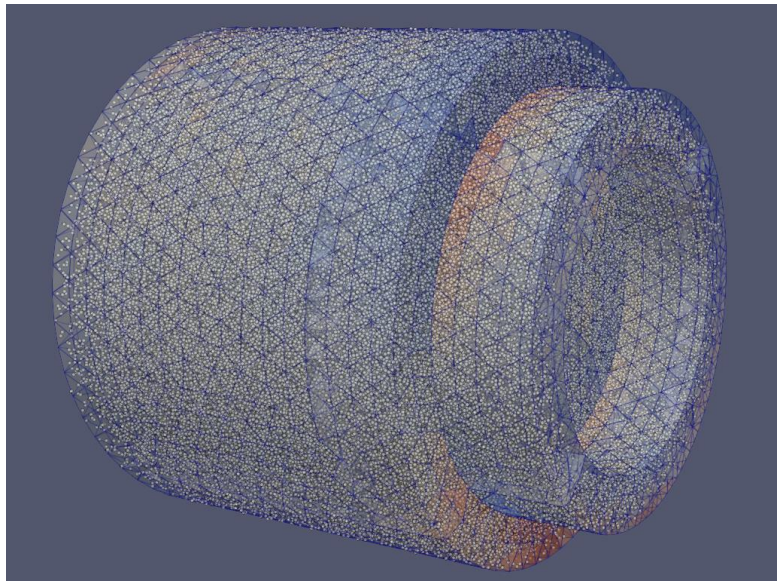


Image from G. Vijaykumar

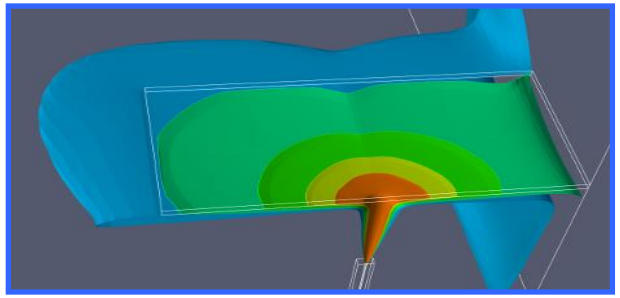
Where will fractures in ice sheets occur?



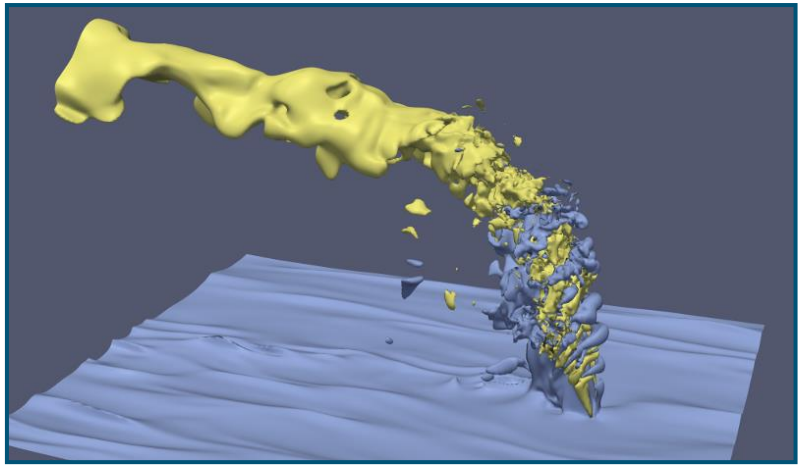
How much charge will accumulate on this surface?



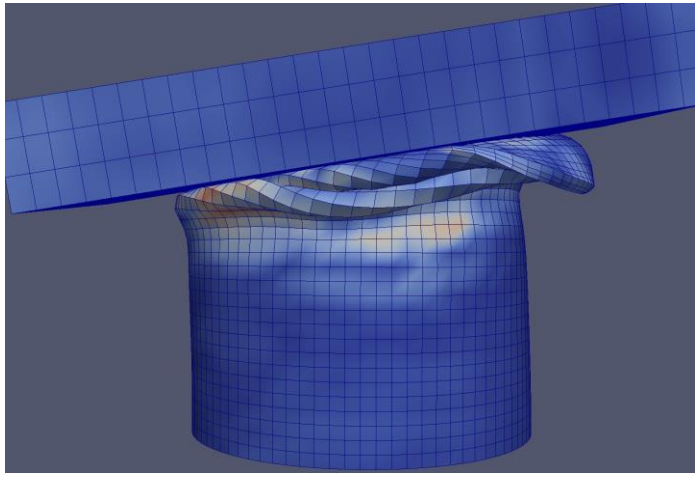
How much hydrogen would leak before detection?



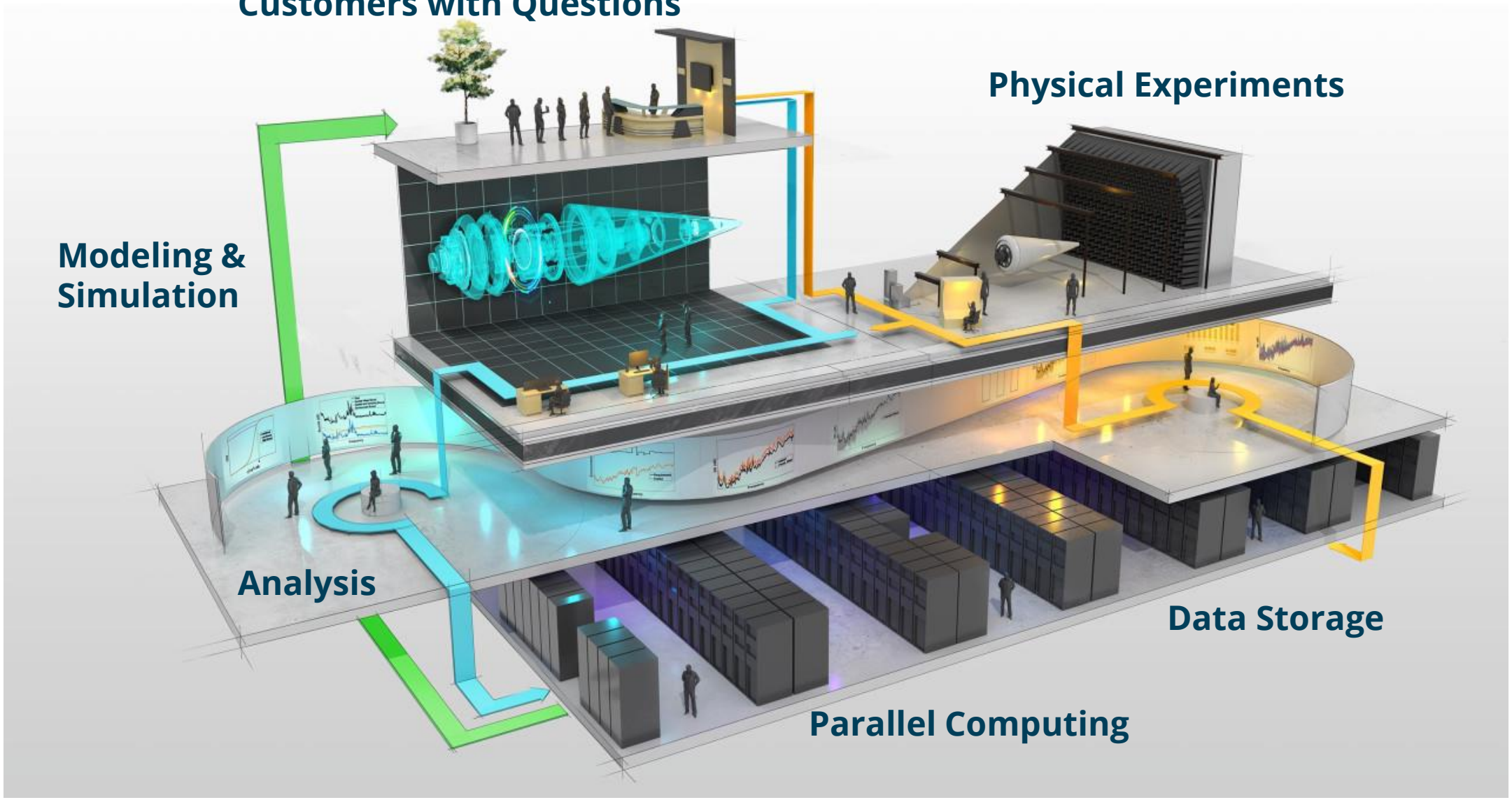
How does mixture ratio affect combustion?



Will this part still function if damaged?



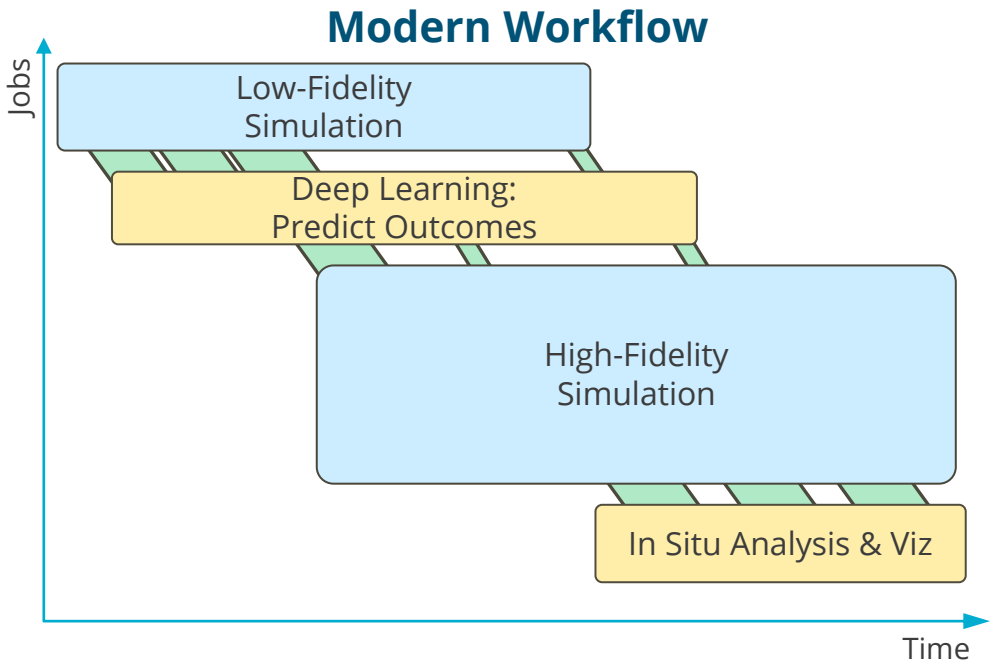
What is the Vision for Computing at the Labs?

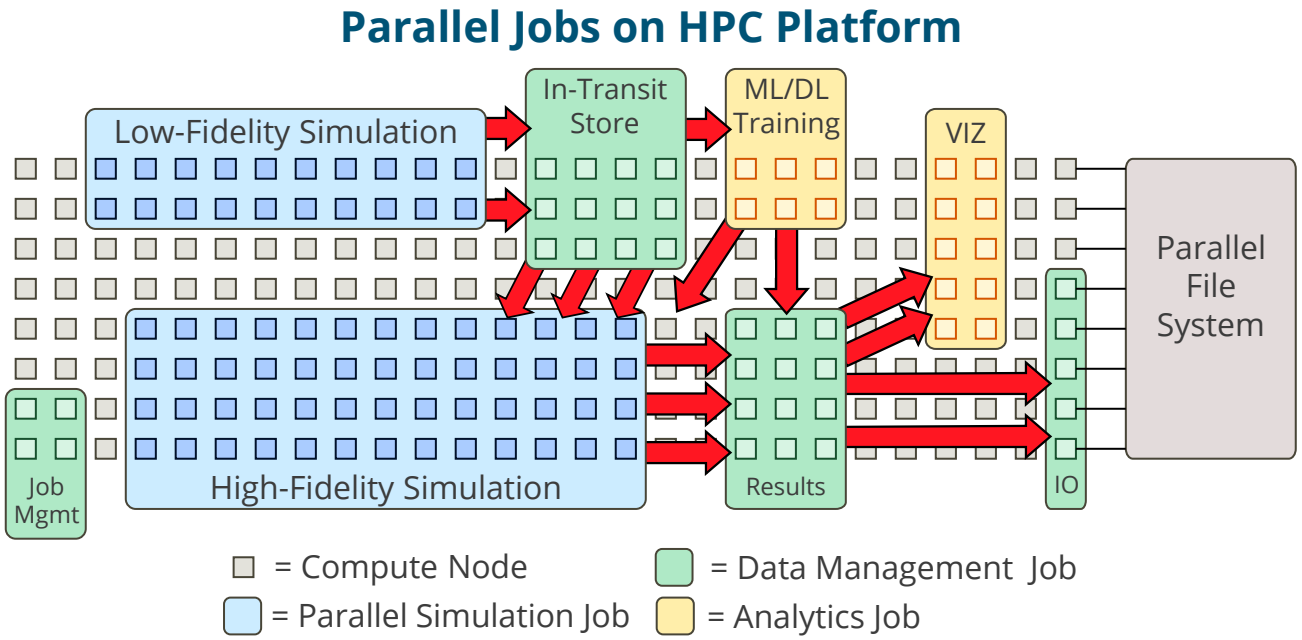


How can SmartNICs improve workflows on HPC Platforms?

High-Performance Computing Workflows

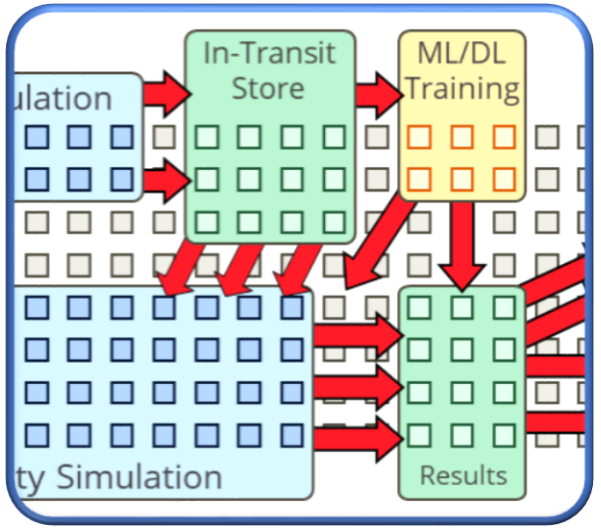
- Scientists run massively-parallel simulations to answer difficult research questions
  - Runtime datasets are too large to store (Summit ~3PB of RAM)
  - Couple different analysis tools to simulations to harvest information
- Modern workflows involve multiple parallel applications





Data Management and Storage Services

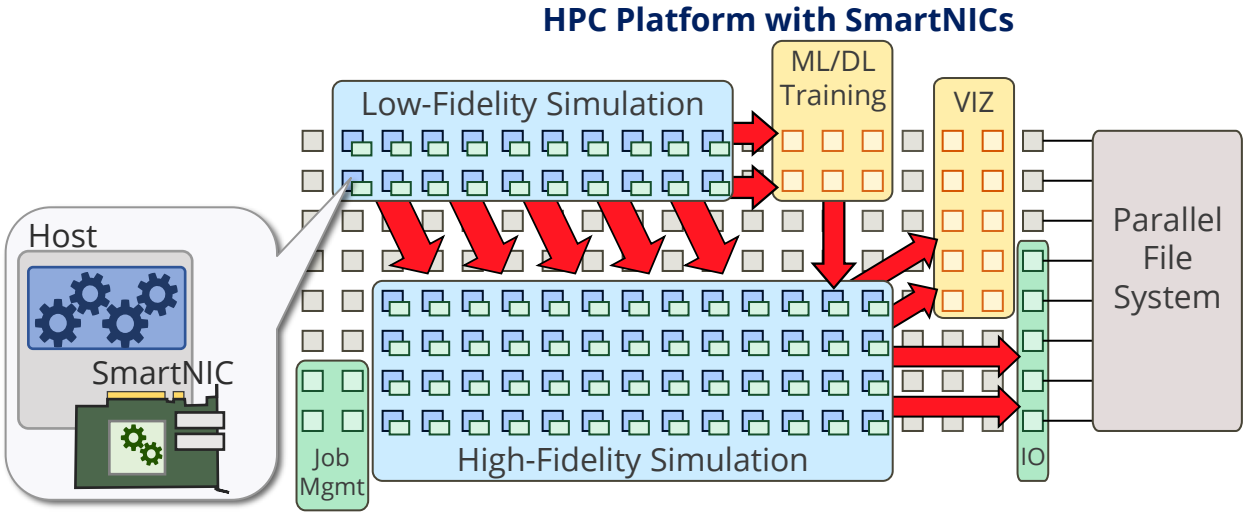
- HPC community has multiple data management libraries for routing data between jobs
  - DataSpaces, Mochi, Conduit, FAODEL
- Distributed memory services
  - Dedicate a number of nodes to serve as a pool for housing objects in memory
  - Use RDMA methods and event-driven semantics to move objects efficiently
- 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?



Smart Network Interface Cards (SmartNICs)

- Network vendors now offer SmartNICs with *user-programmable* resources
  - Examples: NVIDIA BlueField-2 DPU, Intel IPU, and FPGA
  - Embedded processors provide isolated space for caching and processing in-transit data
- Emerging HPC platforms include SmartNICs
  - How do we make an environment for hosting data services in SmartNICs?

BlueField-2 DPU SmartNIC

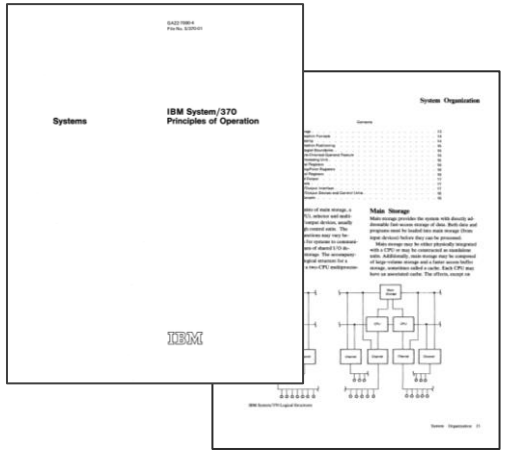


= Compute Node with a *SmartNIC* for offloading data services

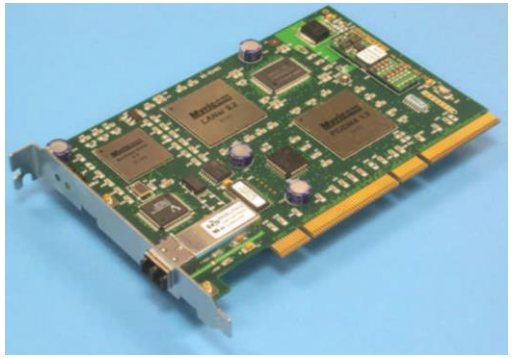
SmartNICs

SmartNIC Timeline

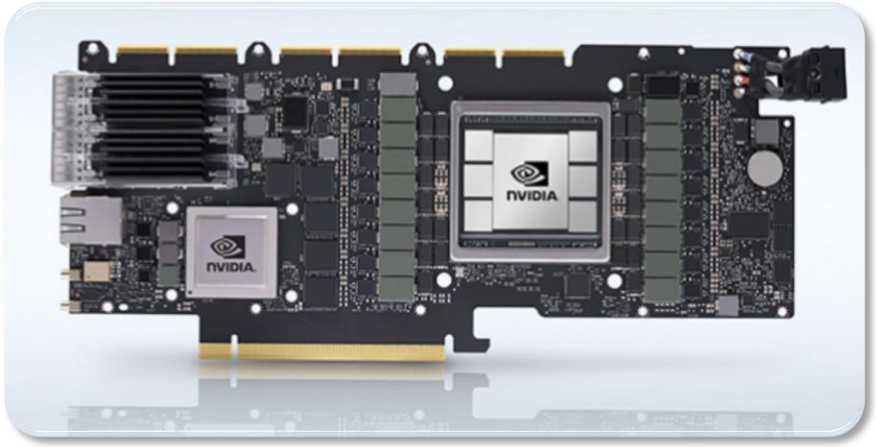
- 1957-1980s: IBM Channel I/O Processors
- 1990s: The Golden Age of Programmable NICs
  - Caltech Mosaic → Myrinet LANai NIC
- 2000s: Queue-based NICs
- 2010s: Security & Data Center NICs
  - Tilera: 64-100 cores for network pattern matching
  - Cloud vendors (AWS, Azure): Secure Networks
- 2016: Mellanox BlueField SmartNICs and DPUs
  - 2020: BlueField-2 Ethernet (InfiniBand in 2021)
  - 2023: BlueField-3: “10x improvement”
  - 202x: BlueField-4: “100x improvement” (GPU)



IBM Channel I/O

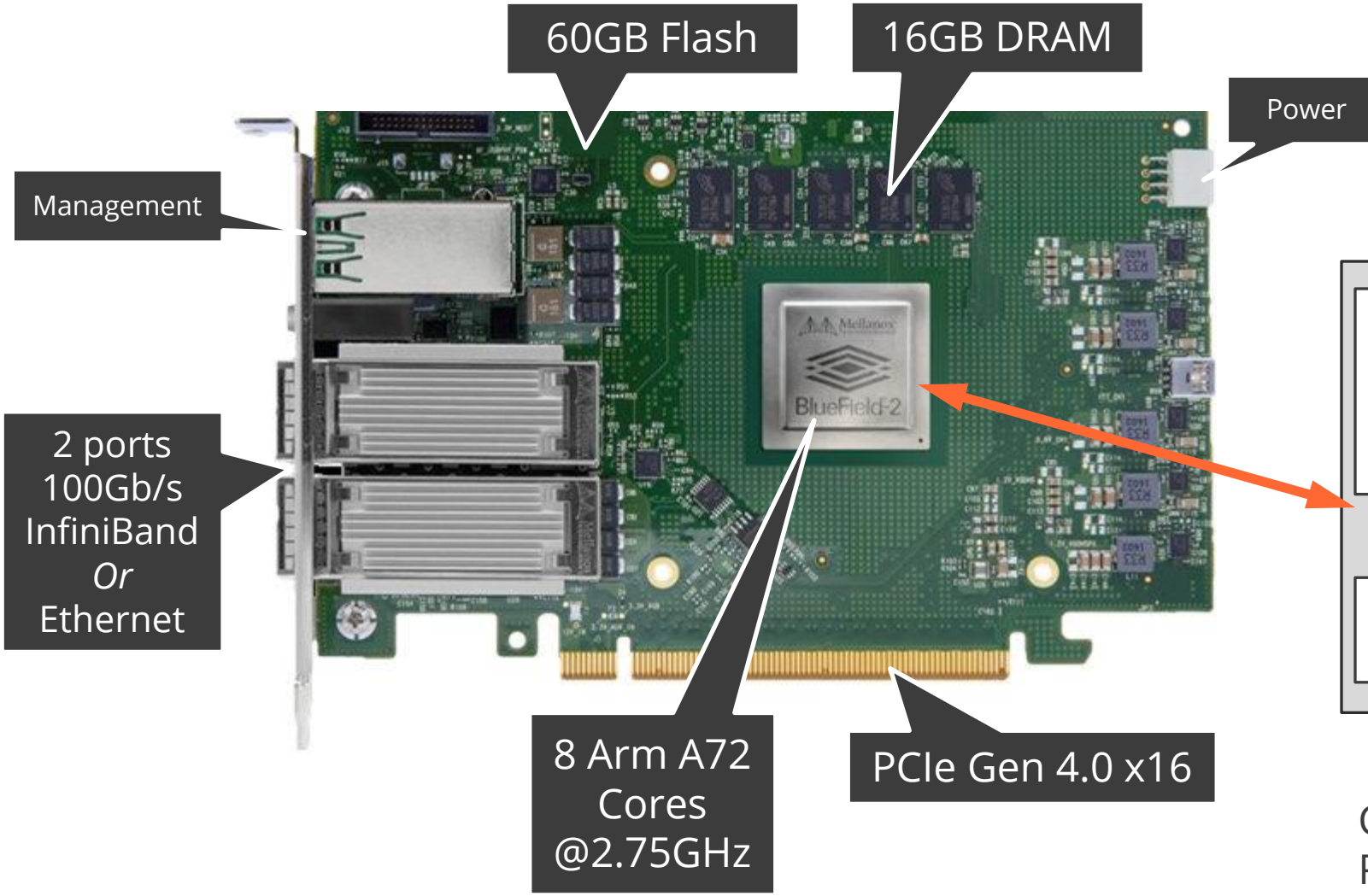


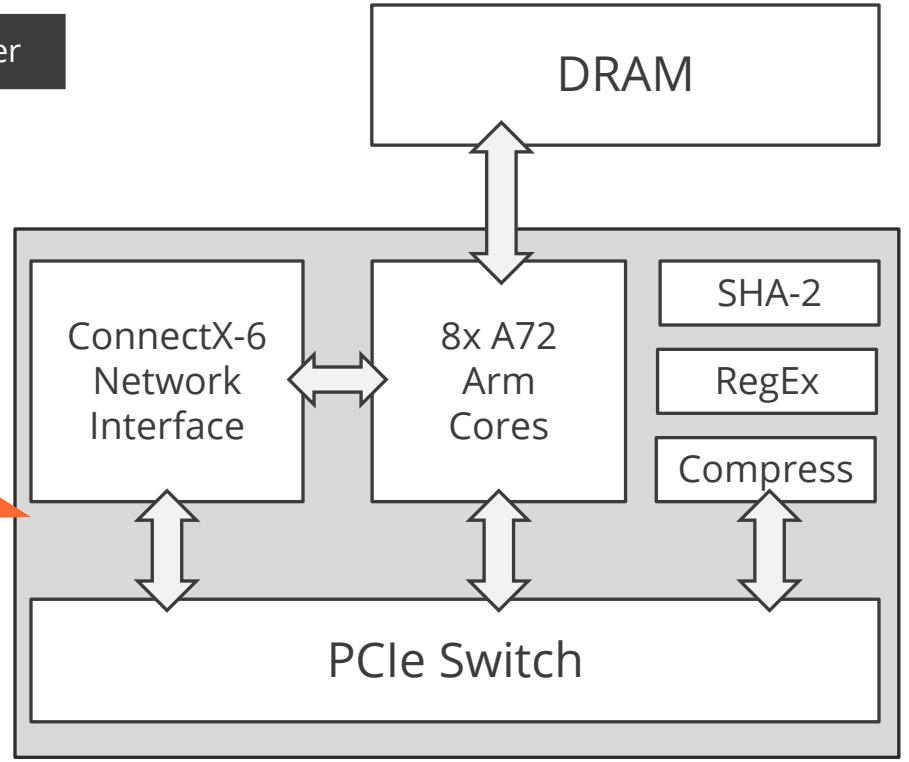
Myricom LANai



NVIDIA Converged

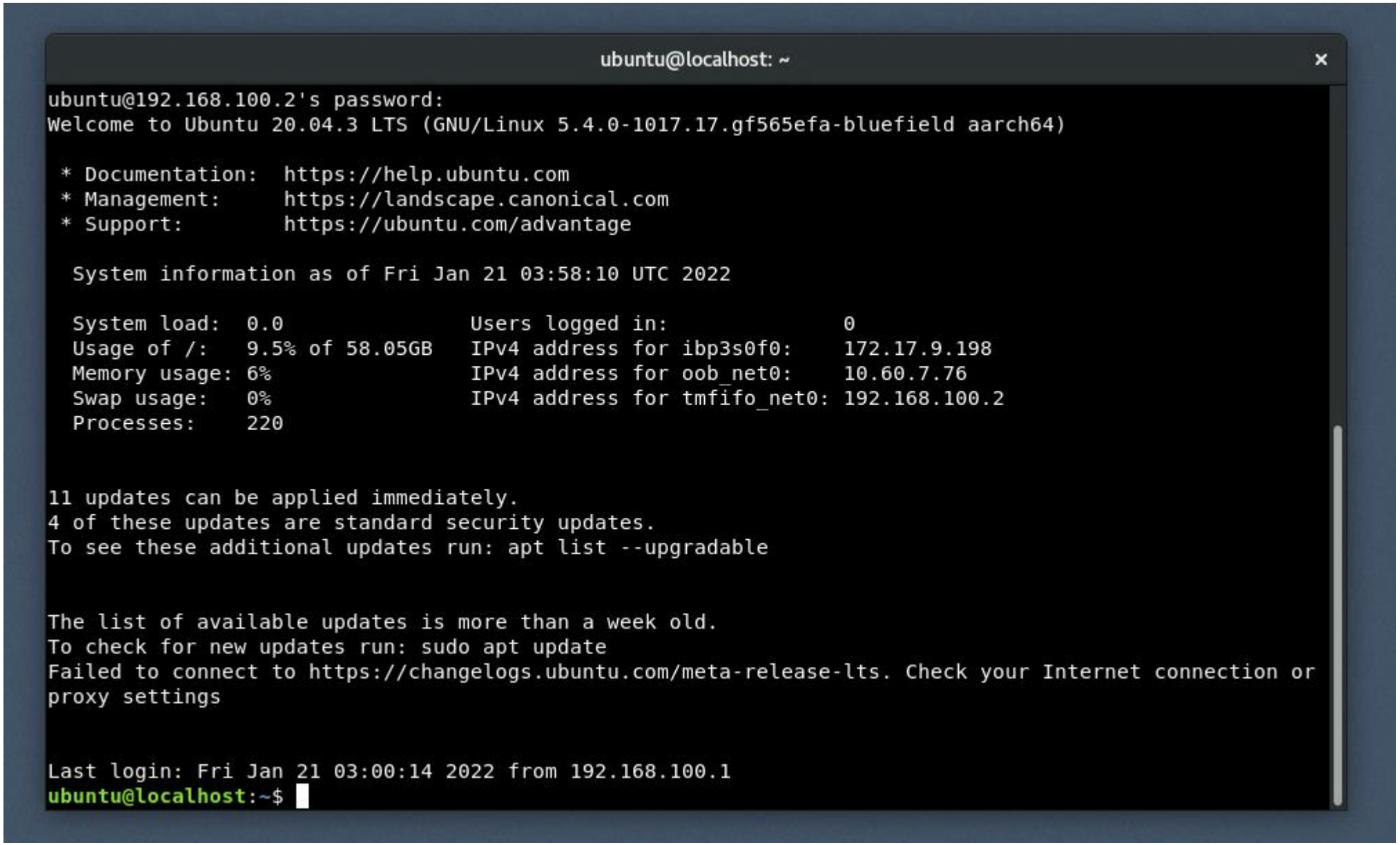
Today: NVIDIA BlueField-2 VPI SmartNICs





Cost: ~$2,000
Power: 30W idle, 42W active, 63W max
OS: Linux (Ubuntu, RHEL, etc)

User Interface: Linux



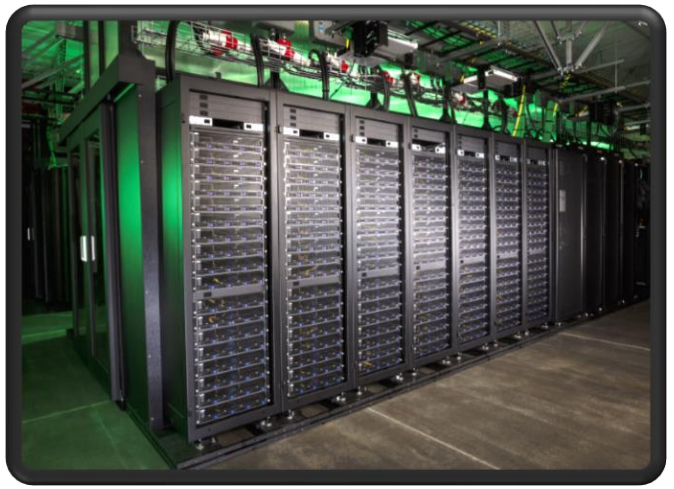
Expect Embedded Processor Performance

- SmartNIC designers face multiple constraints
  - Cost: Niche market, price sensitive
  - PCIe card: limited space, power, and cooling
- Embedded processors: 4-10x slower, but better power efficiency

	SmartNIC	Host CPU	GPU
Processor	Arm A72	AMD EPYC 7543P	Ampere A100
Cores	**8**	**32**	**108 SMs**
Clock	2.75GHz	2.8GHz	0.765 – 1.41GHz
L1 Cache	256KB	1MB	192KB
L2 Cache	6MB	256MB	-
Memory Capacity	16GB	512GB	40GB
Memory Bandwidth	**25GB/s**	**204GB/s**	**1,555GB/s**
TDP	63W	225W	250W

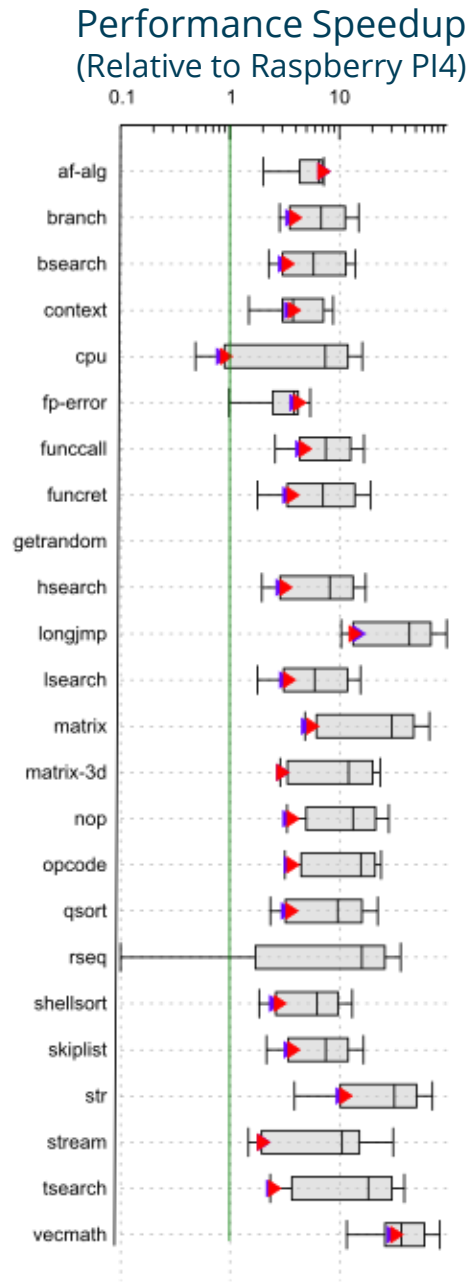
Sandia's Glinda Cluster (2021)

- 126 Compute Nodes
- Zen3, BlueField-2, A100



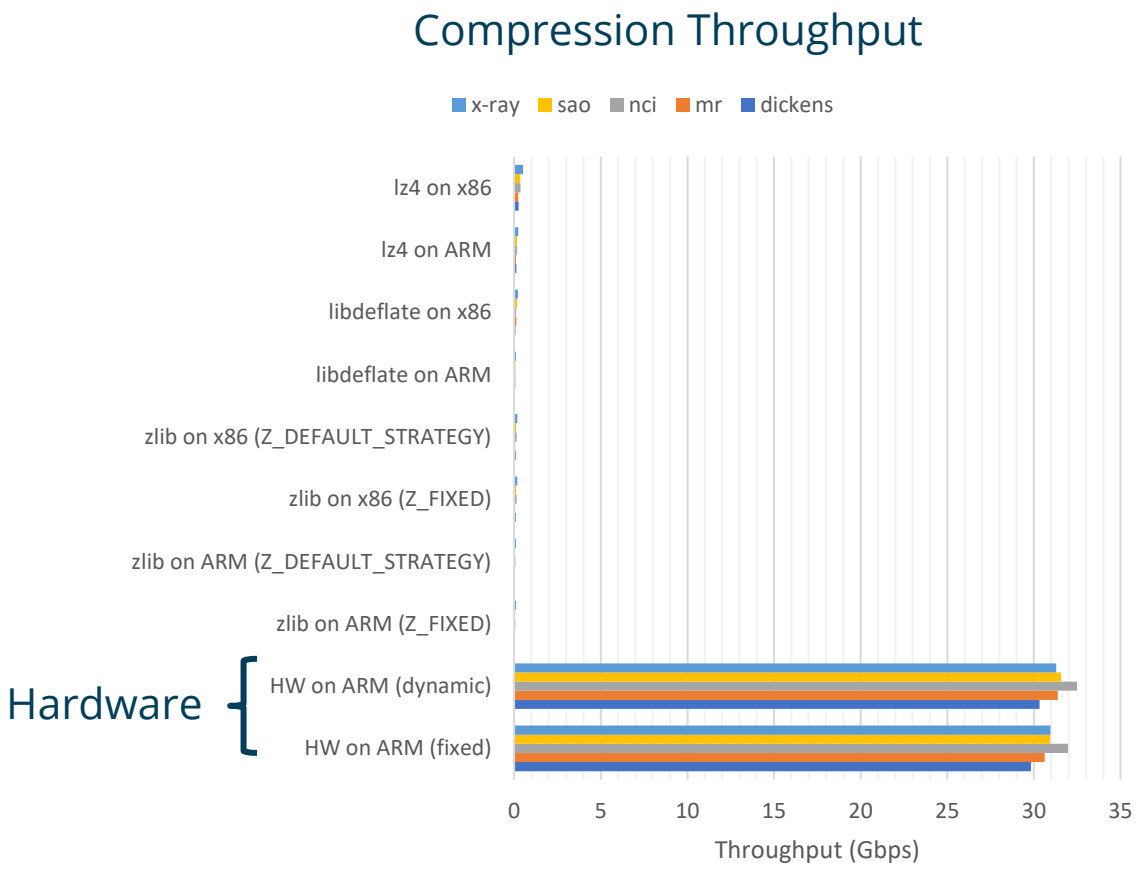
Microbenchmarks

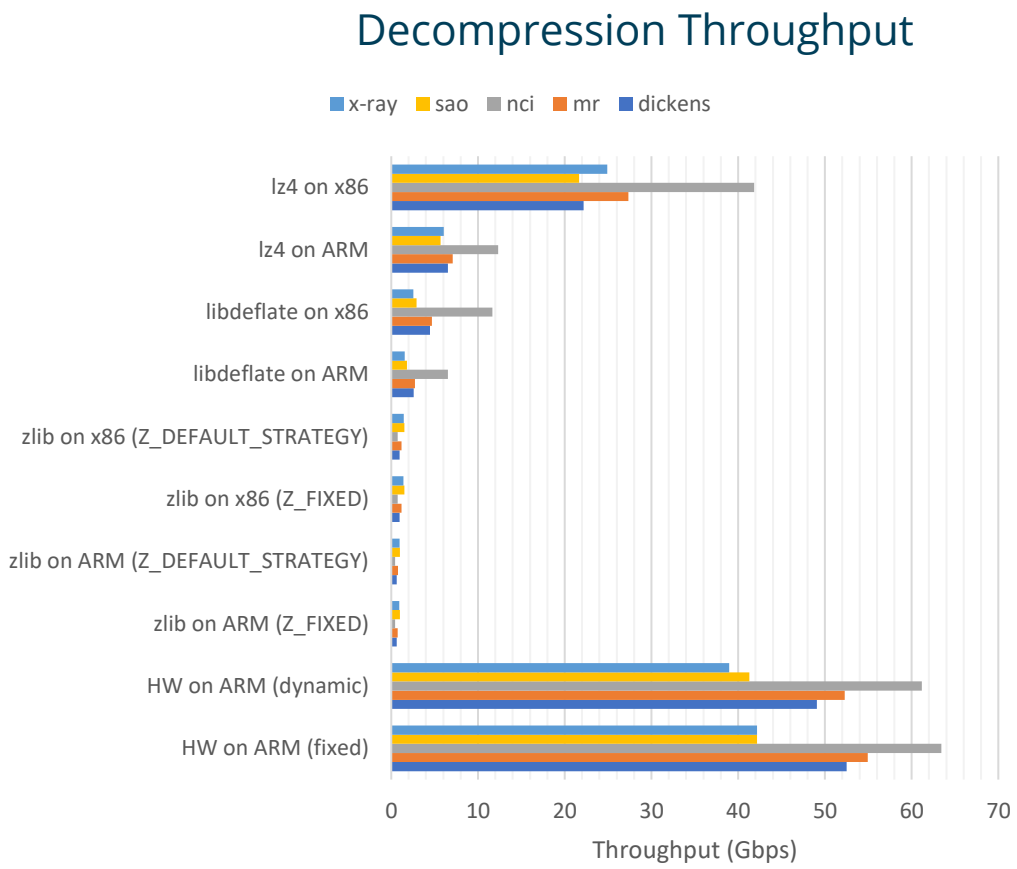
- **stress-ng** from Colin Ian King (Canonical)
  - Suite of 218 stressors for kernel burn-in
  - Normalize results to Raspberry Pi 4B
- Comparison of BlueField-2 to 12 servers
  - Performance range of different servers
  - BlueField-2 2.50 GHz (Ethernet)
  - BlueField-2 2.75 GHz (Ethernet/InfiniBand)
- BlueField-2's Arms order of magnitude slower than hosts
  - *...but still good enough for data management tasks*



BlueField-2's Compression is Significantly Faster than Host

- BlueField-2 features compression hardware for the DEFLATE algorithm (e.g., gzip)

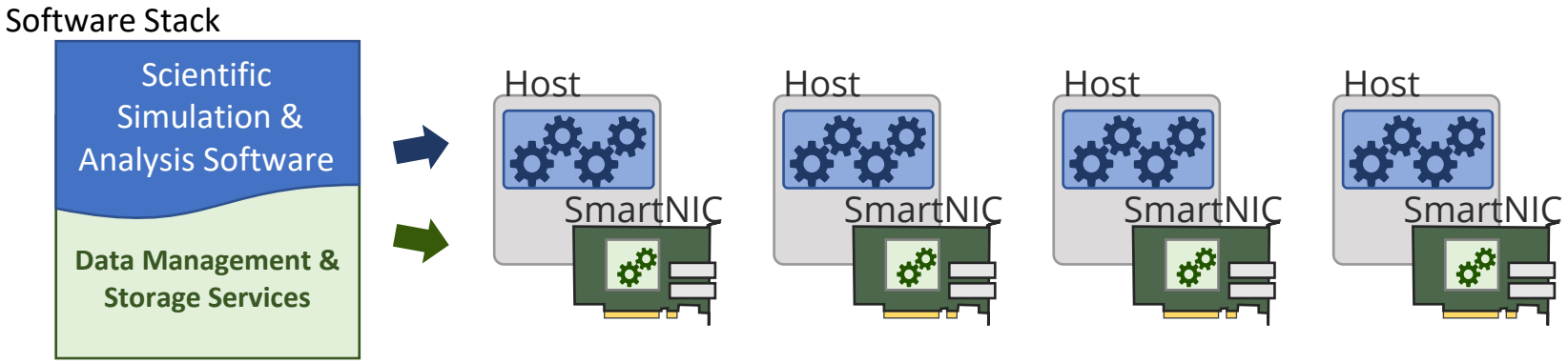




Creating an Environment for Data Services on SmartNICs

Create an Environment for Hosting Data Services on SmartNICs

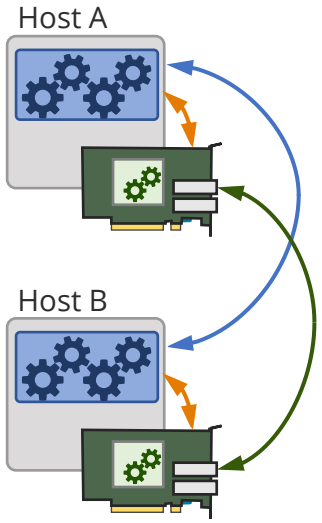
- We define five requirements (R1-R5) for creating this environment
  - Three communication, Two computation
- Leverage existing libraries as much as possible
- Prototype environment
  - Communication via **Faodel**: C++ library with distributed-memory Key/Blob API built on RDMA
  - Computation via **Apache Arrow**: C++ library for processing in-memory tabular data



Resolving Communication Requirements with Faodel

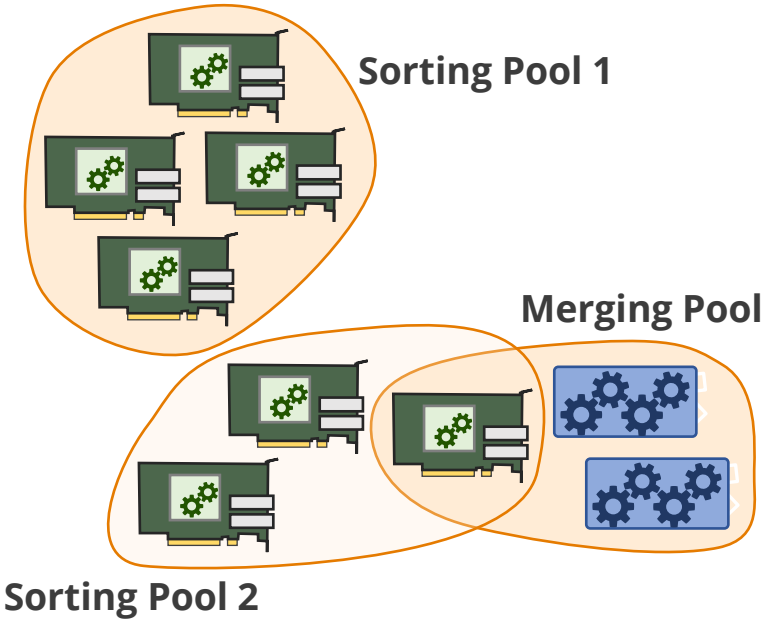
R1: Any-to-Any Transfers

- Faodel has globally accessible endpoints
- Host and SmartNICs can be endpoints
- Put/Get remote objects
- RDMA for point-to-point transfers



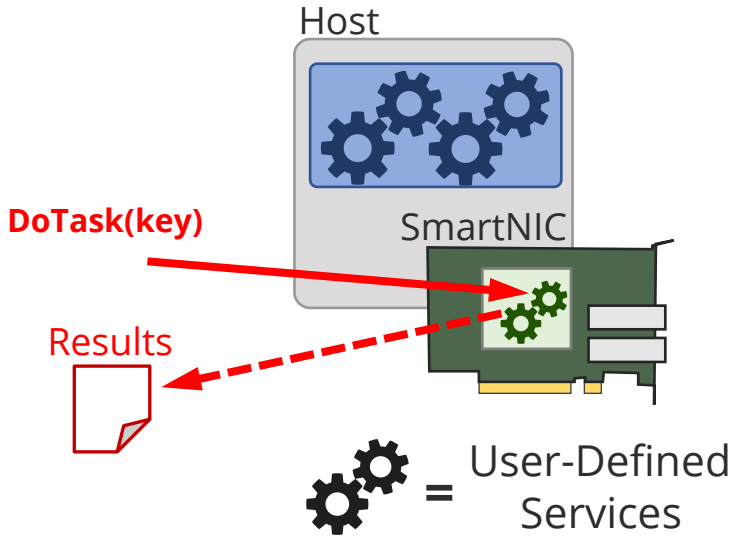
R2: Group Resources

- Faodel uses a Pool abstraction
- Pool has endpoints & distribution policy
- Work mapped to resources at run time



R3: Dispatch Computations

- Faodel primarily moves data
- Invoke remote operation on object
- Local main can also make decisions



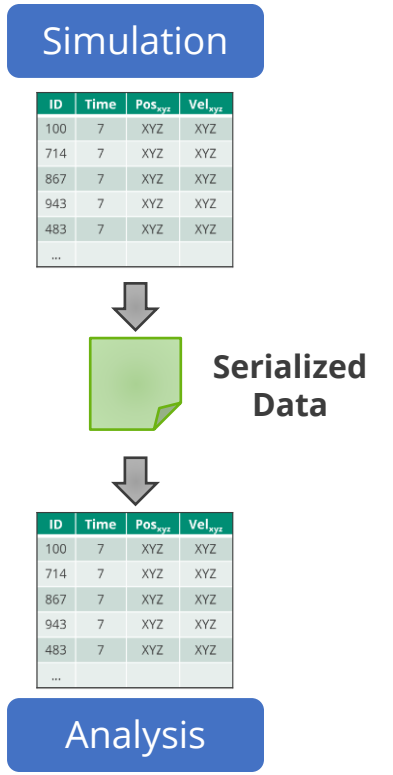
Resolving Computational Requirements with Apache Arrow

R4: Common Data Representation

- Arrow provides robust data structures for 2D data
- Efficient in-memory storage
- Built-in functions to serialize

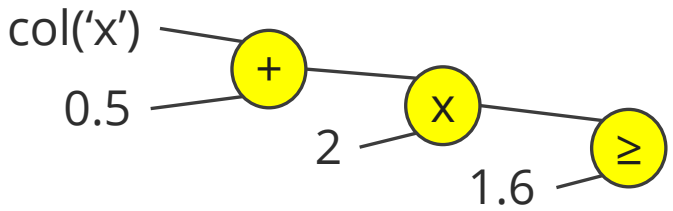
APACHE ARROW

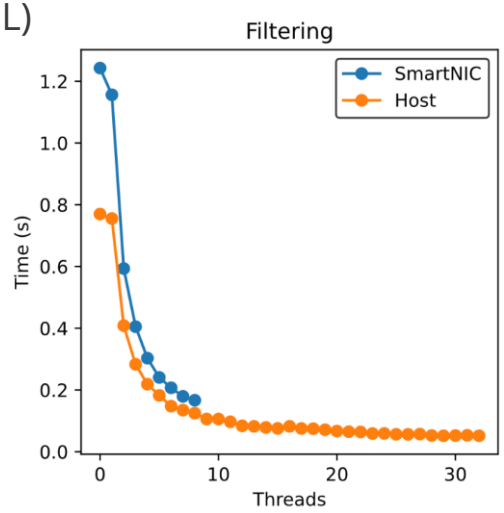
ID	Time	Pos_{xyz}	Vel_{xyz}
100	7	XYZ	XYZ
714	7	XYZ	XYZ
867	7	XYZ	XYZ
943	7	XYZ	XYZ
483	7	XYZ	XYZ
…			



R5: Data-Parallel Computations

- Arrow includes compute functions for tables
- Target for higher-level languages (SQL)
- Thread- and SIMD-Aware



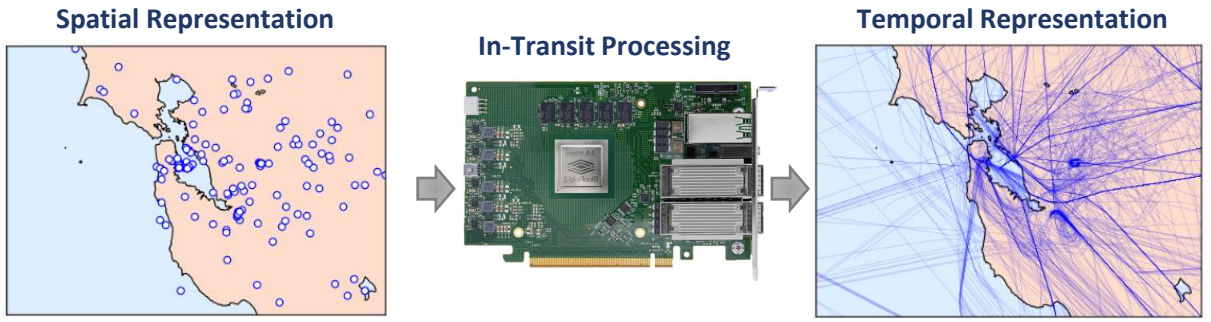



```
// 2 * (0.5 + x) >= 1.6
auto filter_expression = arrow::compute::greater_equal(
    arrow::compute::call(
        "multiply",
        {arrow::compute::literal(2),
         arrow::compute::call("add_checked", {arrow::compute::literal(0.5),
                                              arrow::compute::field_ref("x")})}),
    arrow::compute::literal(1.6));
```

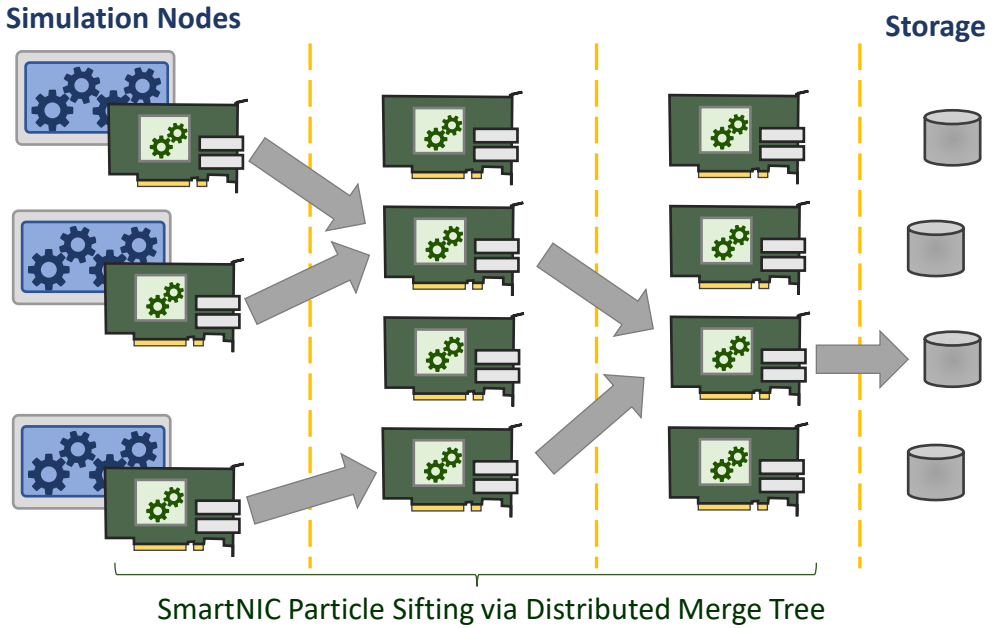
Particle Sifting Example

Example: Reorganizing Particle Simulation Results

- Particle simulations track billions of particles
- Mismatch between producers/consumers
  - Simulations: Sorted by position and time
  - Analytics: Sorted by ID and time
- Particle sifting service
  - Periodically sample current data
  - Use distributed SmartNICs to reorganize
  - Distributed merge tree sorts data by ID
- Implementation
  - Faodel Pools/Keys to control data flow
  - Arrow compute to split data
- Experiments on 100-node Cluster w/ BlueField-2 DPUs

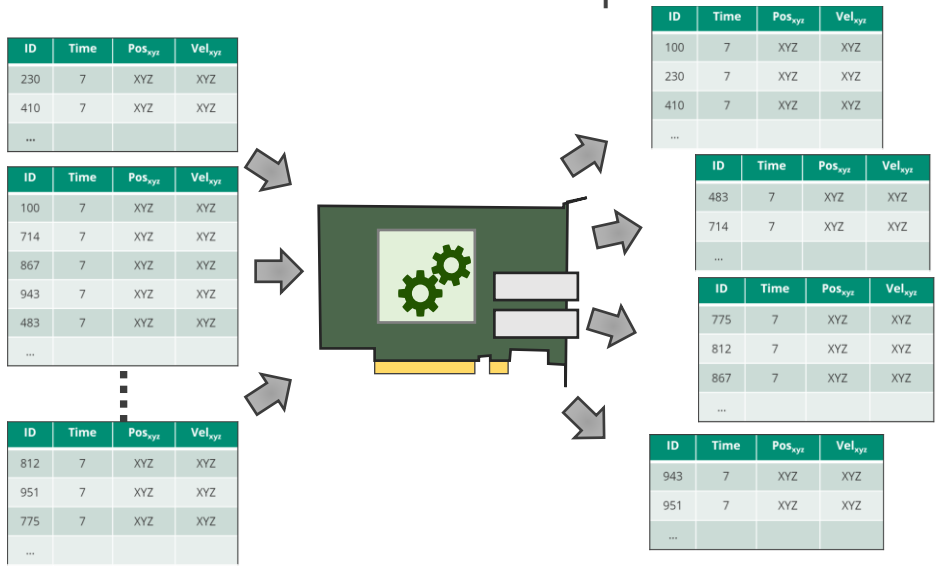


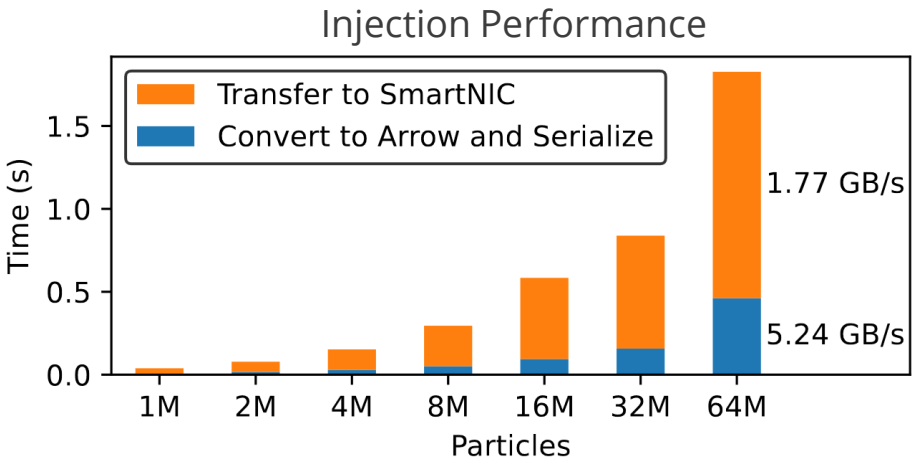
SmartNICs enable simulation results to be transformed while in transit to storage.

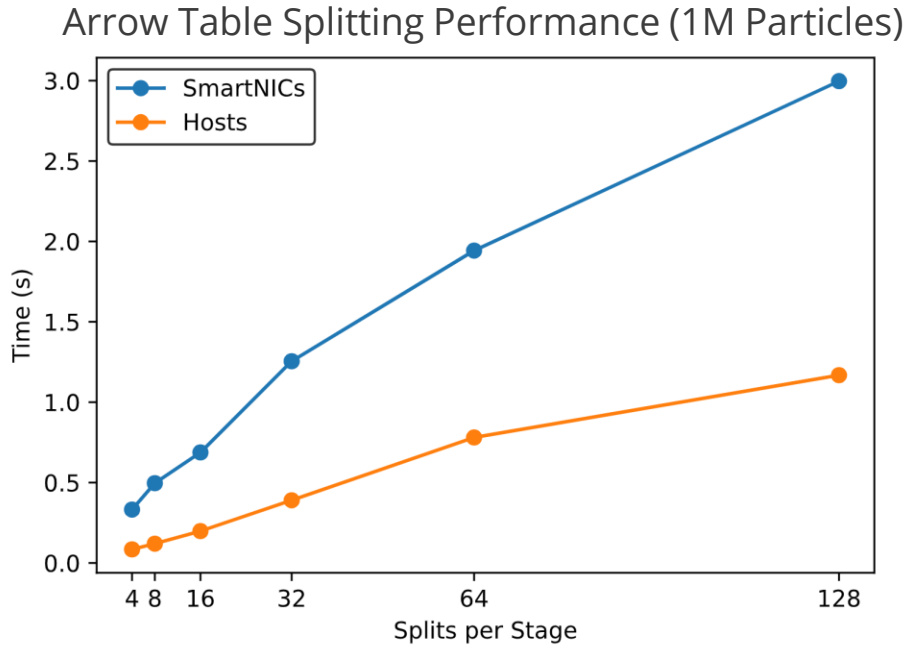


Performance Measurements

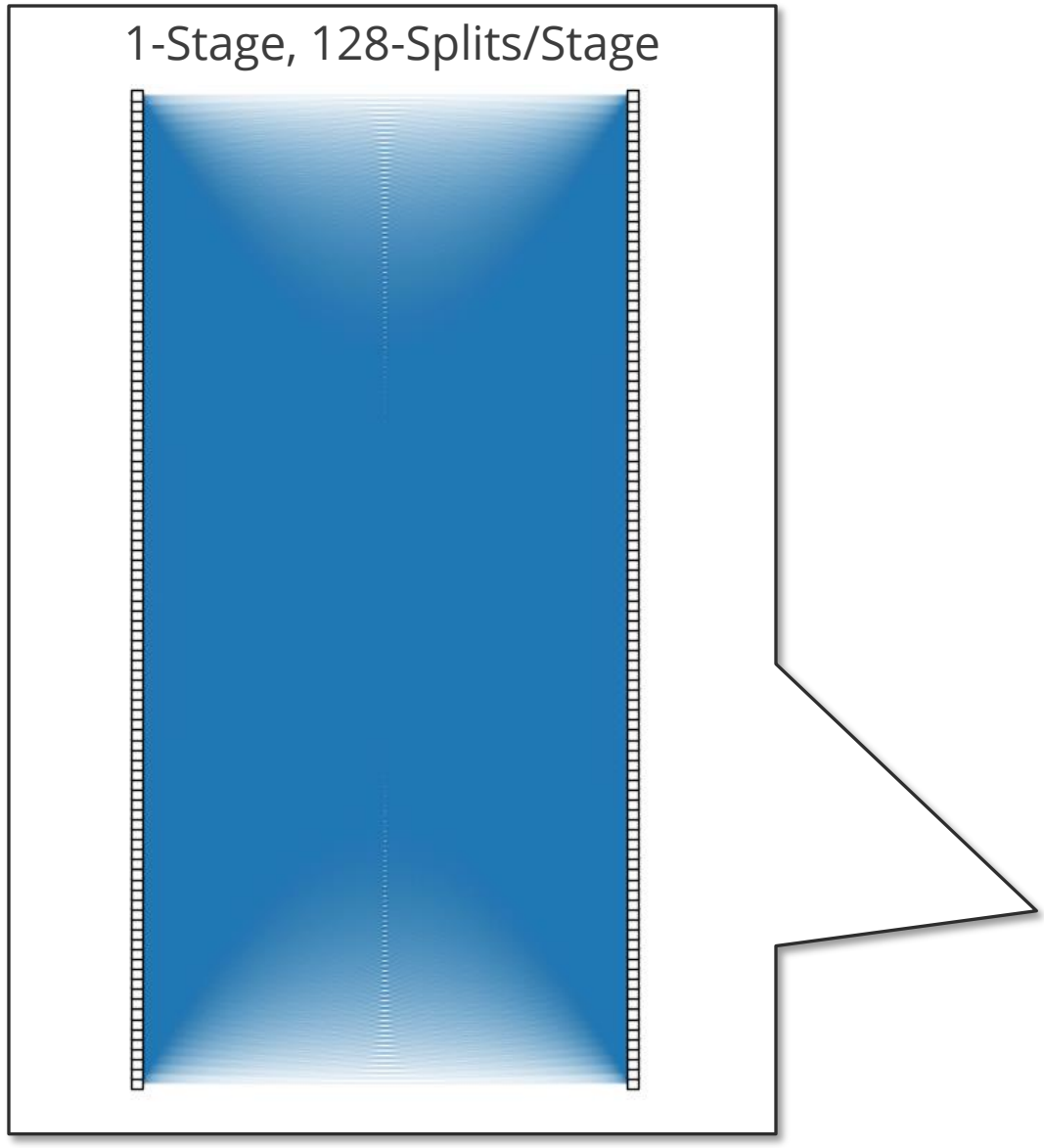
- Injection
  - Convert to Apache Arrow's serialized IPC format
  - Transfer to local SmartNIC
  - 1M–64M Particles (37MB-2.4GB), Overall: 1.32GB/s
  - ***New Work***: 10GB/s for "Serialize-on-Transfer"
- Splitting Tables
  - Merge incoming tables and split based on particle IDs
  - Implemented with Arrow Compute function

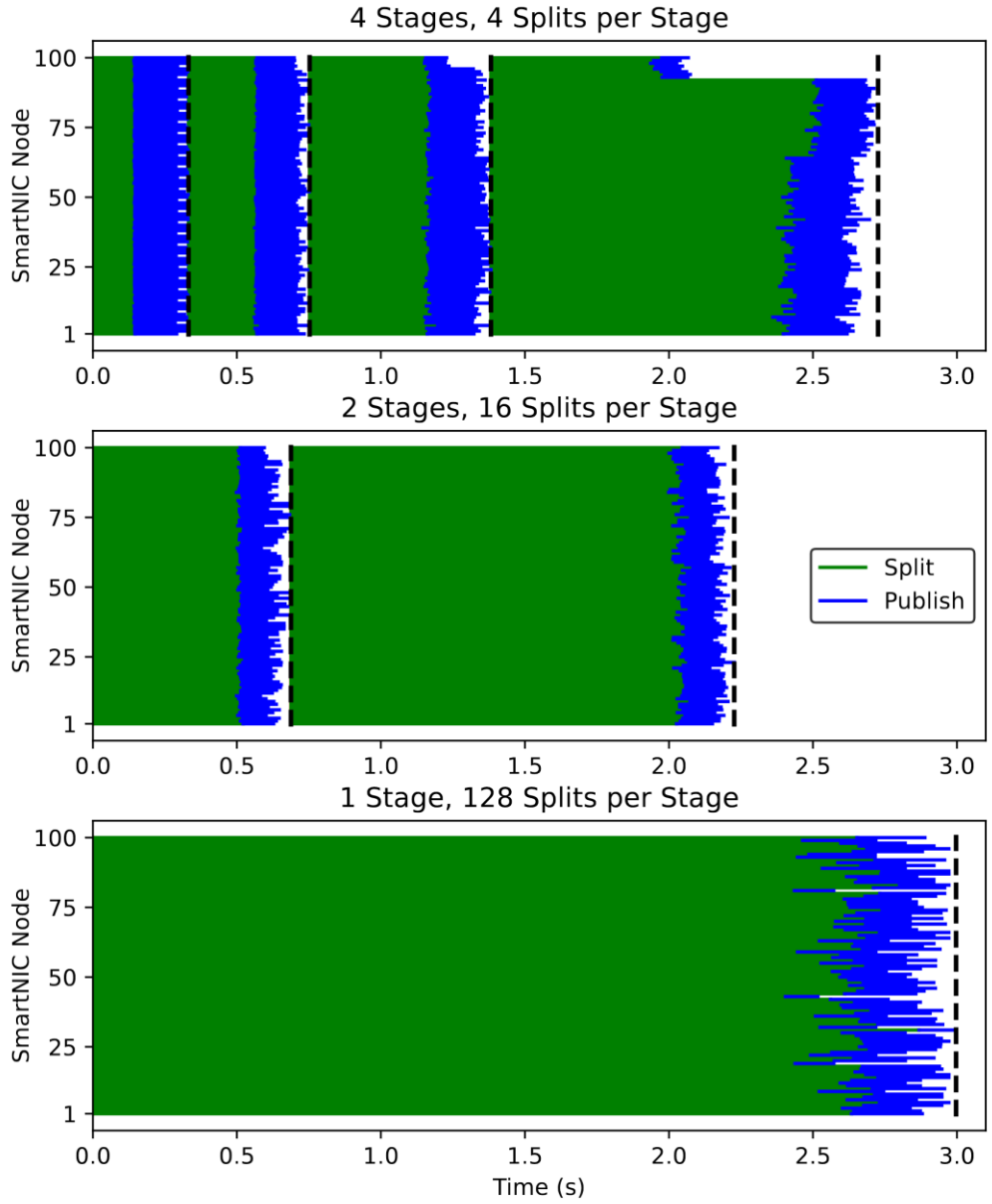




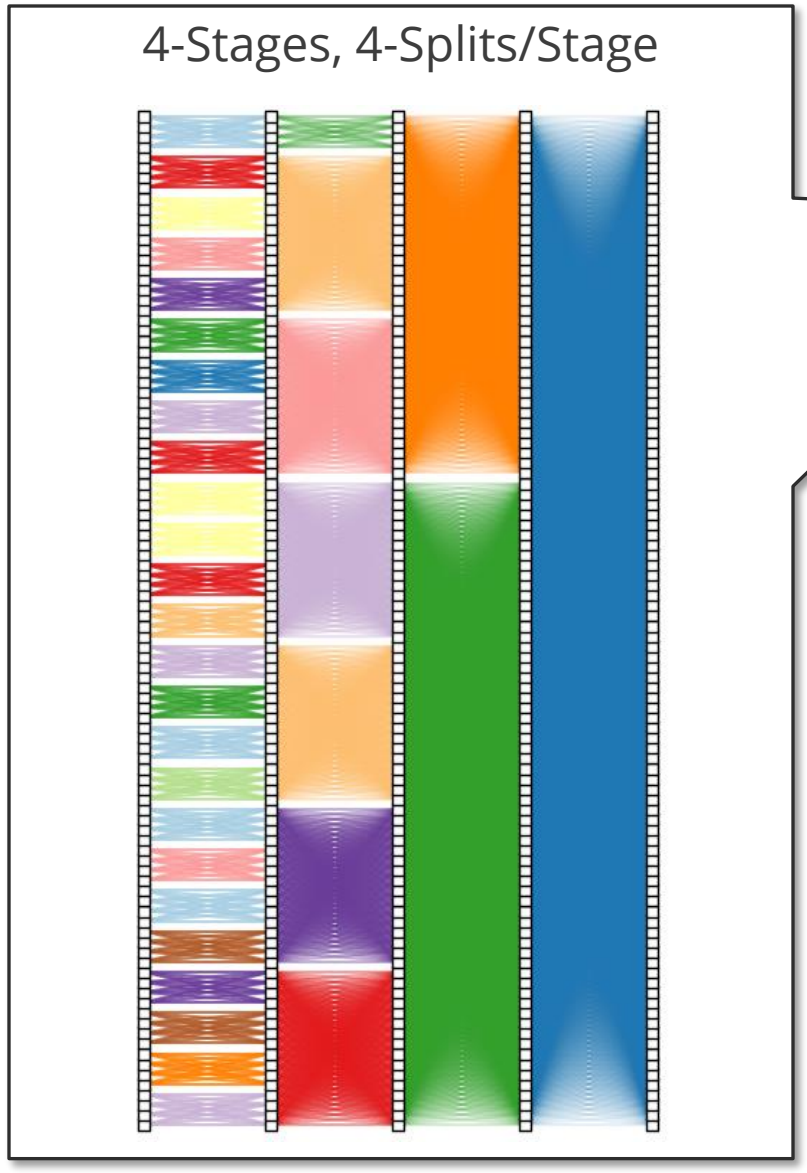


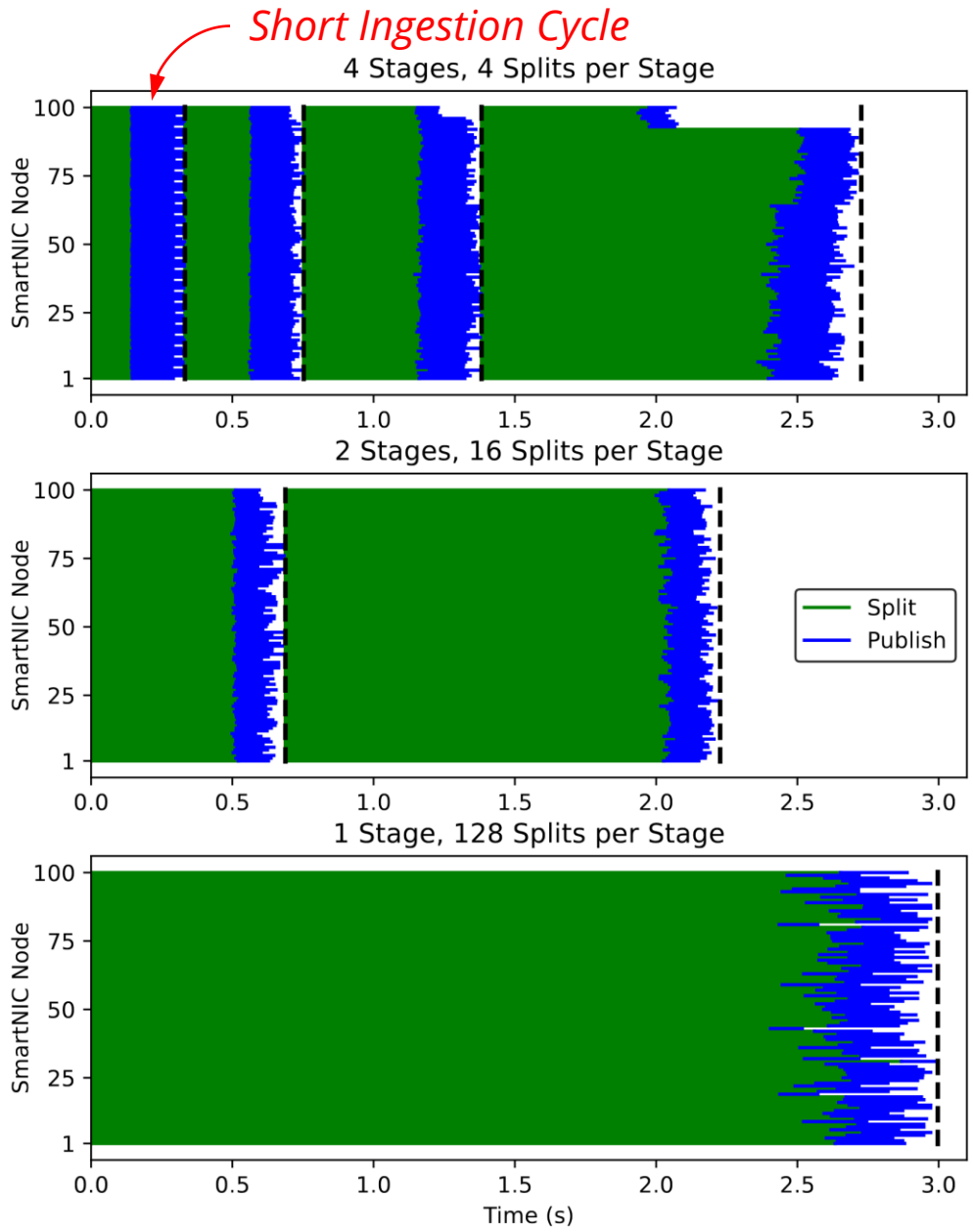
Overall Sifting Performance: 100M Particles on 100 SmartNICs



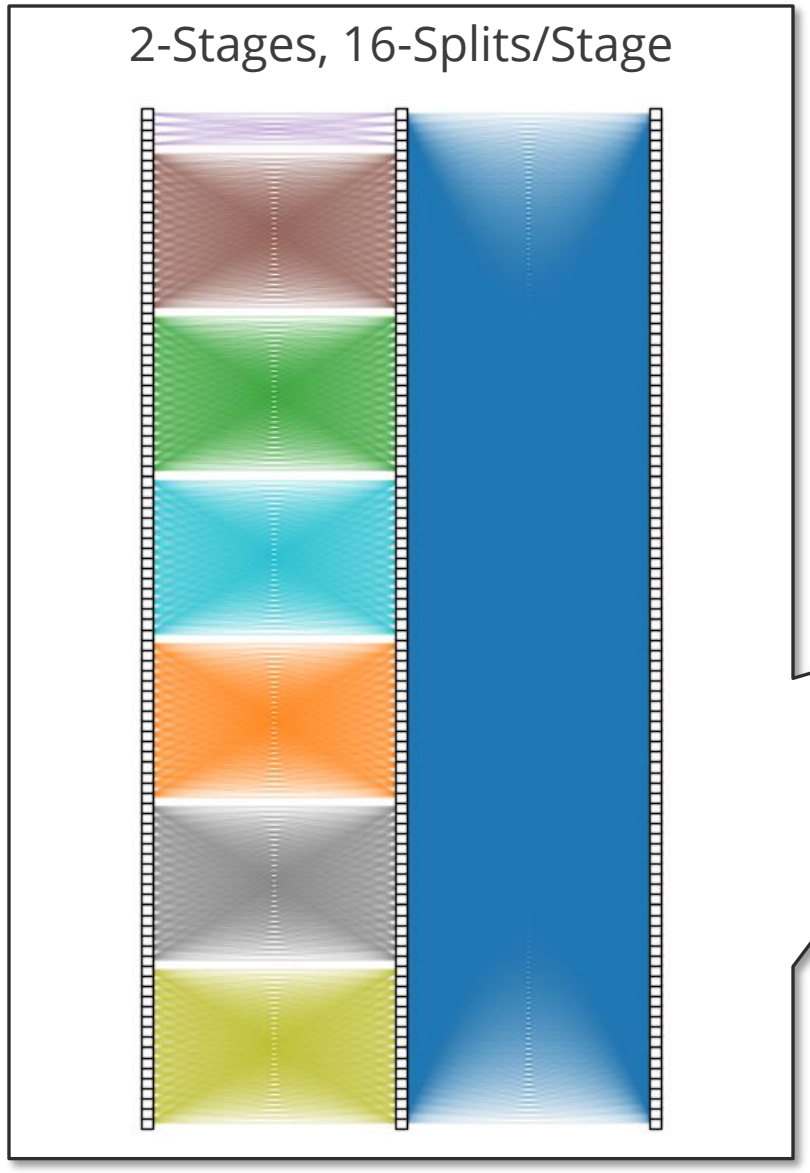


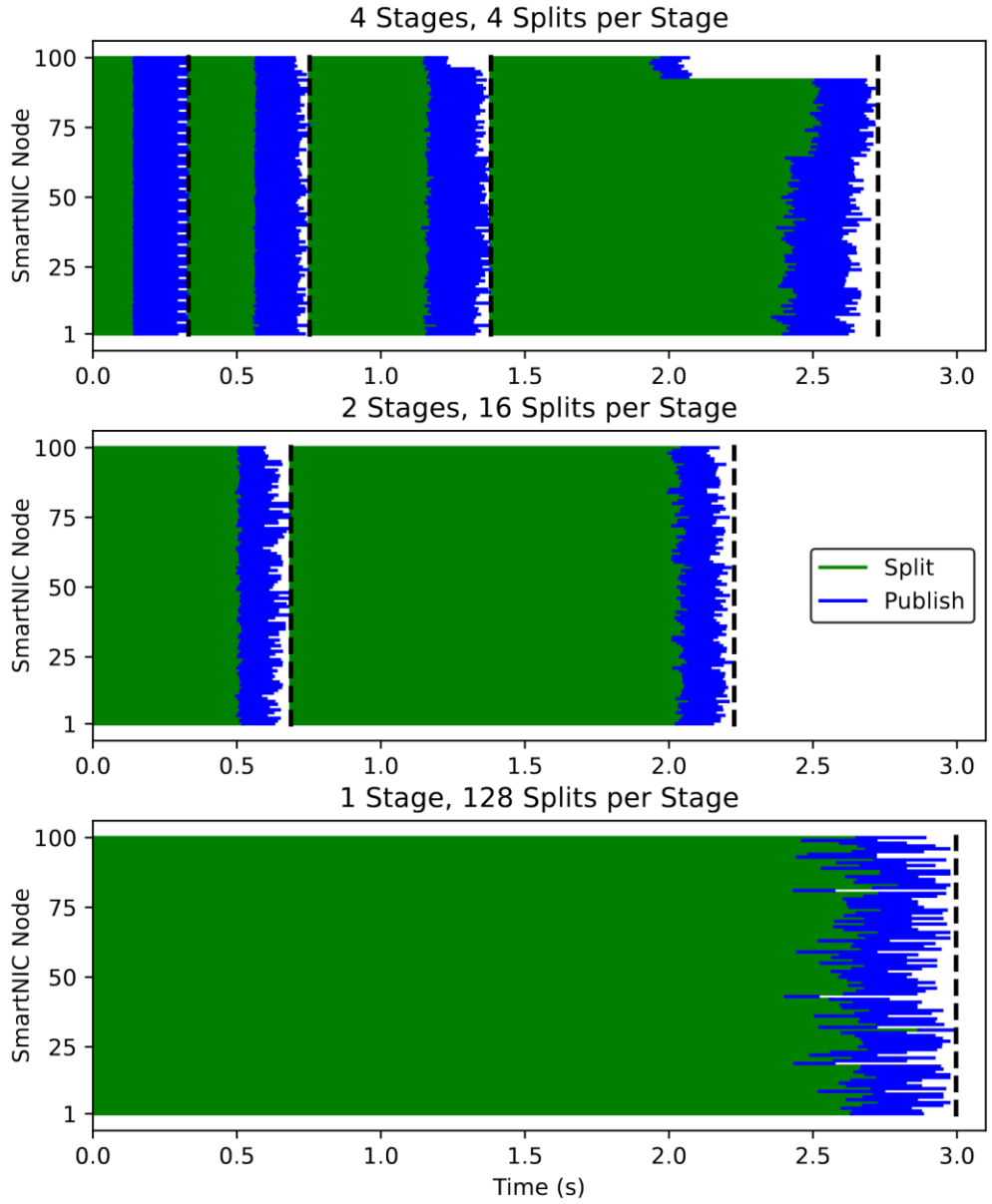
Overall Sifting Performance: 100M Particles on 100 SmartNICs





Overall Sifting Performance: 100M Particles on 100 SmartNICs



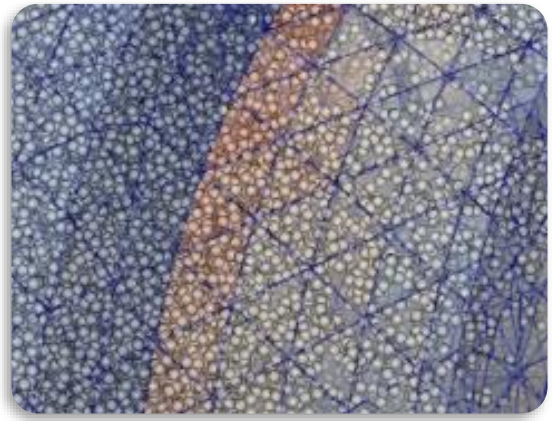


Summary and Future Work

- SmartNICs offer a new space for hosting data management services
  - Positive: Isolated space for operations near producers
  - Negative: Host processors 4x faster, Vendor-specific libraries, extra costs ($, power)
- Can build a functional environment for hosting services from existing libraries
  - Faodel and Arrow provided primitives we needed
- Recent work: Implemented query service for in-transit data
- Future Work
  - Evaluating BlueField-3 and competing SmartNICs
  - Connect with computational storage devices
  - Extend work to other workflows

https://github.com/sandialabs/faodel

https://github.com/apache/arrow



Jobs at the National Labs

- **Students**: The national labs have large, summer jobs programs (paid)
  - Apply in December/January
  - https://www.energy.gov/jobs-national-labs
  - https://www.sandia.gov/careers/
- **Faculty**: Student interns are a great way to develop collaborations with the labs