OAK RIDGE National Laboratory

Streaming Data in HPC Workflows Using ADIOS

Norbert Podhorszki (ORNL)

G. Eisenhauer, N. Podhorszki, A. Gainaru, S. Klasky, P. Davis, M. Parashar

M. Wolf, E. Suchtya, E. Fredj, V. Bolea, F. Pöschel, K. Steiniger, M. Bussmann, R. Pausch, S. Chandrasekaran

Cray Users Group Meeting

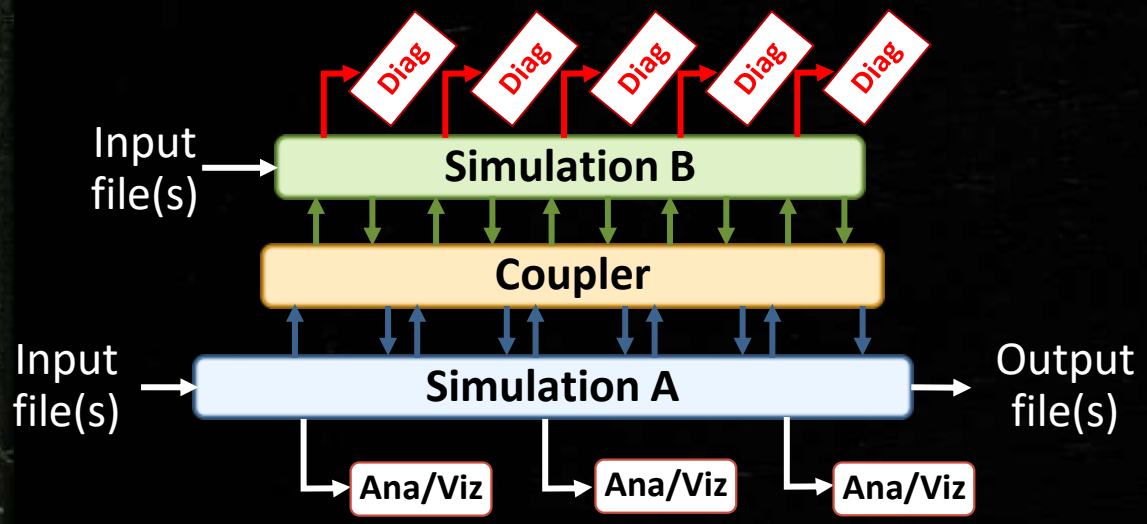
May 2024

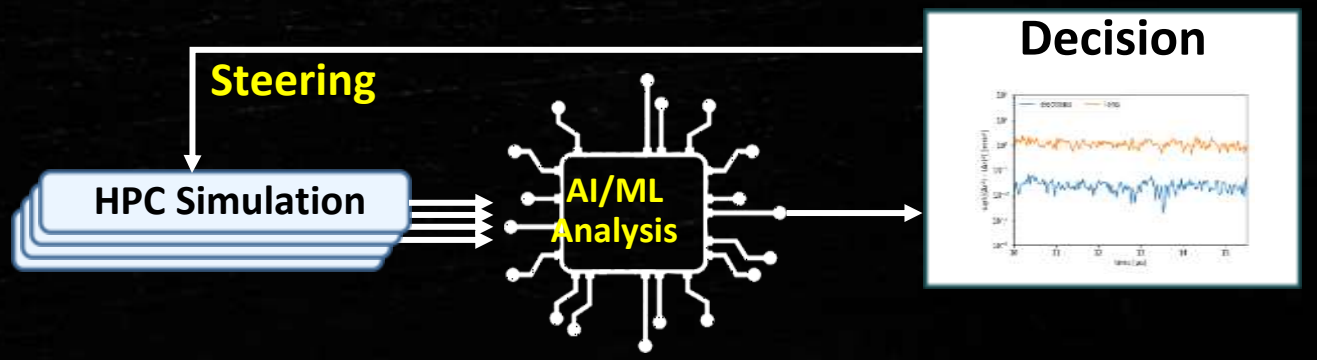
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

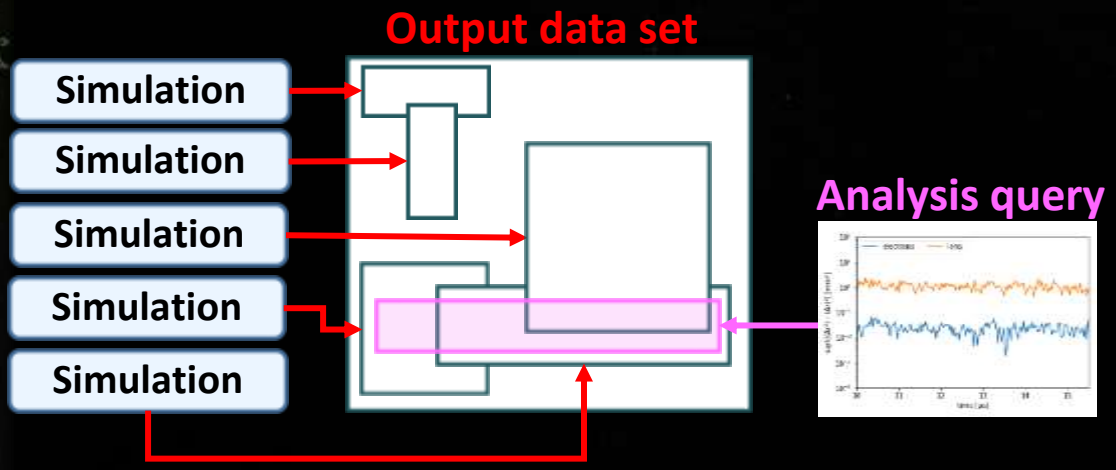
U.S. DEPARTMENT OF ENERGY

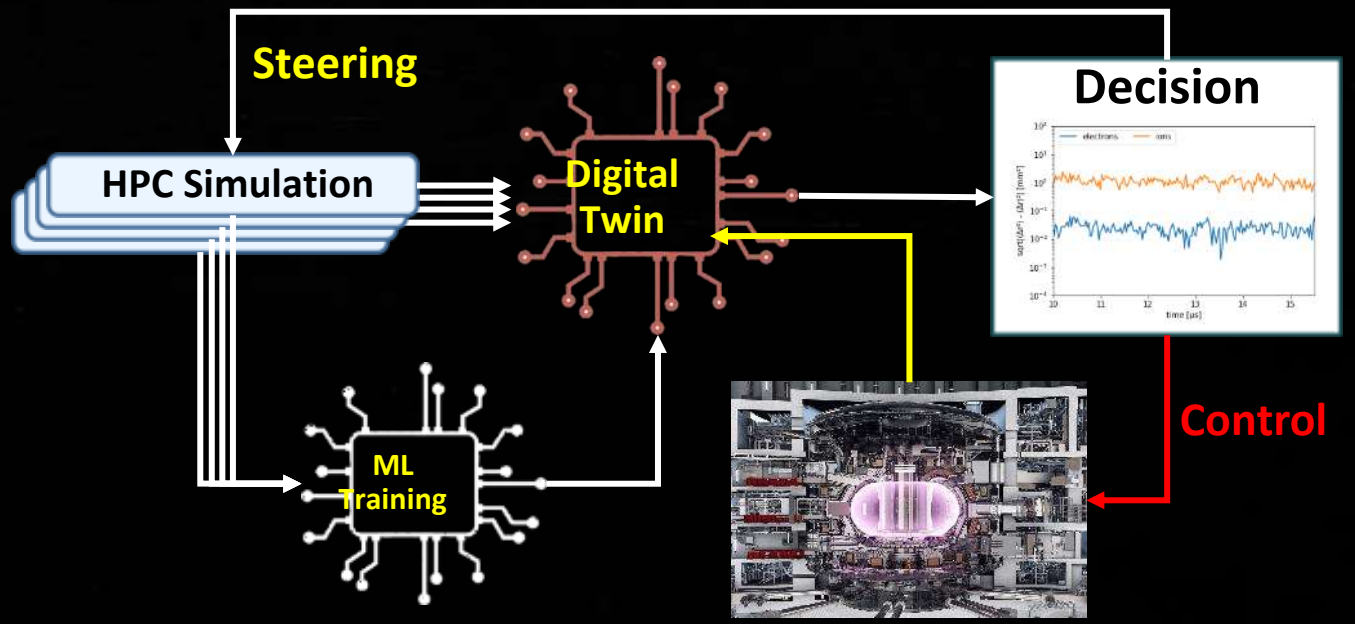
CEDA HIFISTELL

Moving towards next generation workflows







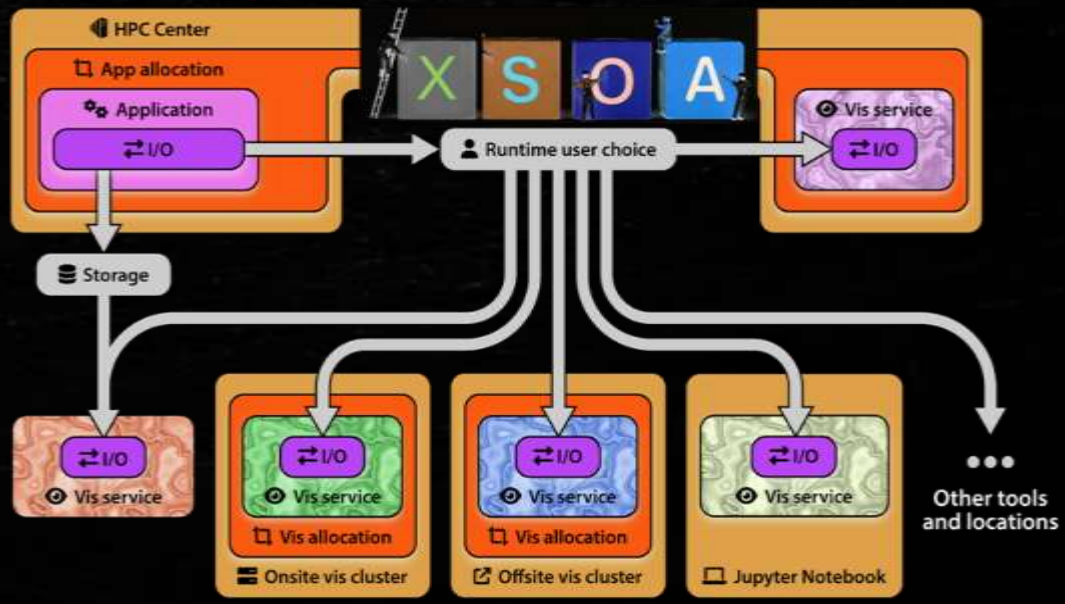


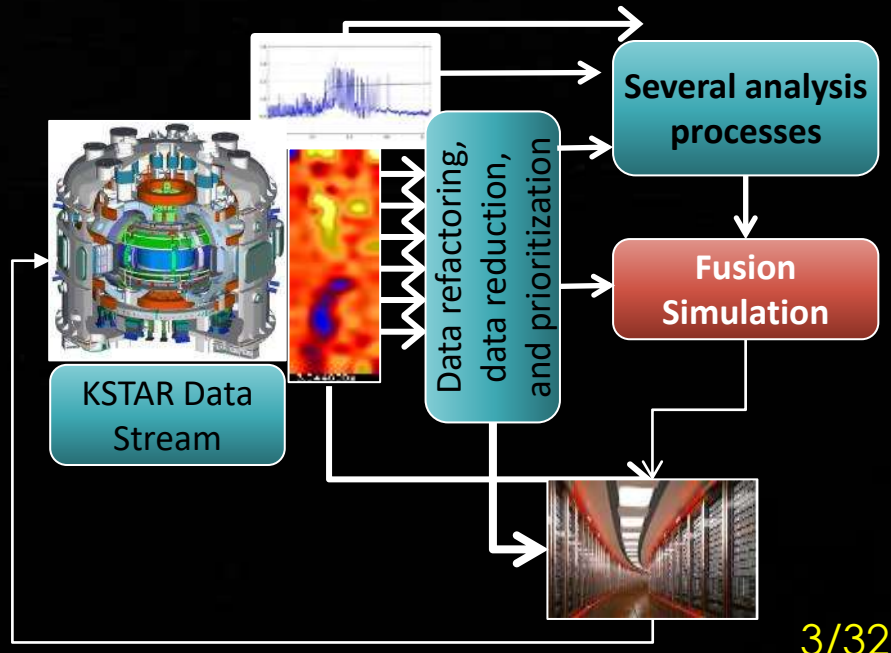
Next Generation workflows need explicit data movement & command/control

Our vision: creating a pub/sub system for high performance SDM

ADIOS allows applications to publish and subscribe to data

- With no modifications, any code can tap into the I/O system for in-line, in-situ, in-transit, WAN, storage system changes
- Data can stream in a “refactored” and filtered/queried in a manner allowing the “most important” information to be prioritized
- Data written and read to storage will be highly optimized on HPC resources and queryable in a federated system





D. Pugmire, J. Huang, S. Klasky, and K. Moreland: The Need for Pervasive In Situ Analysis and Visualization (P-ISAV), WOIV’22.

pnorbert@ornl.gov

Sustainable Staging Transport (SST)

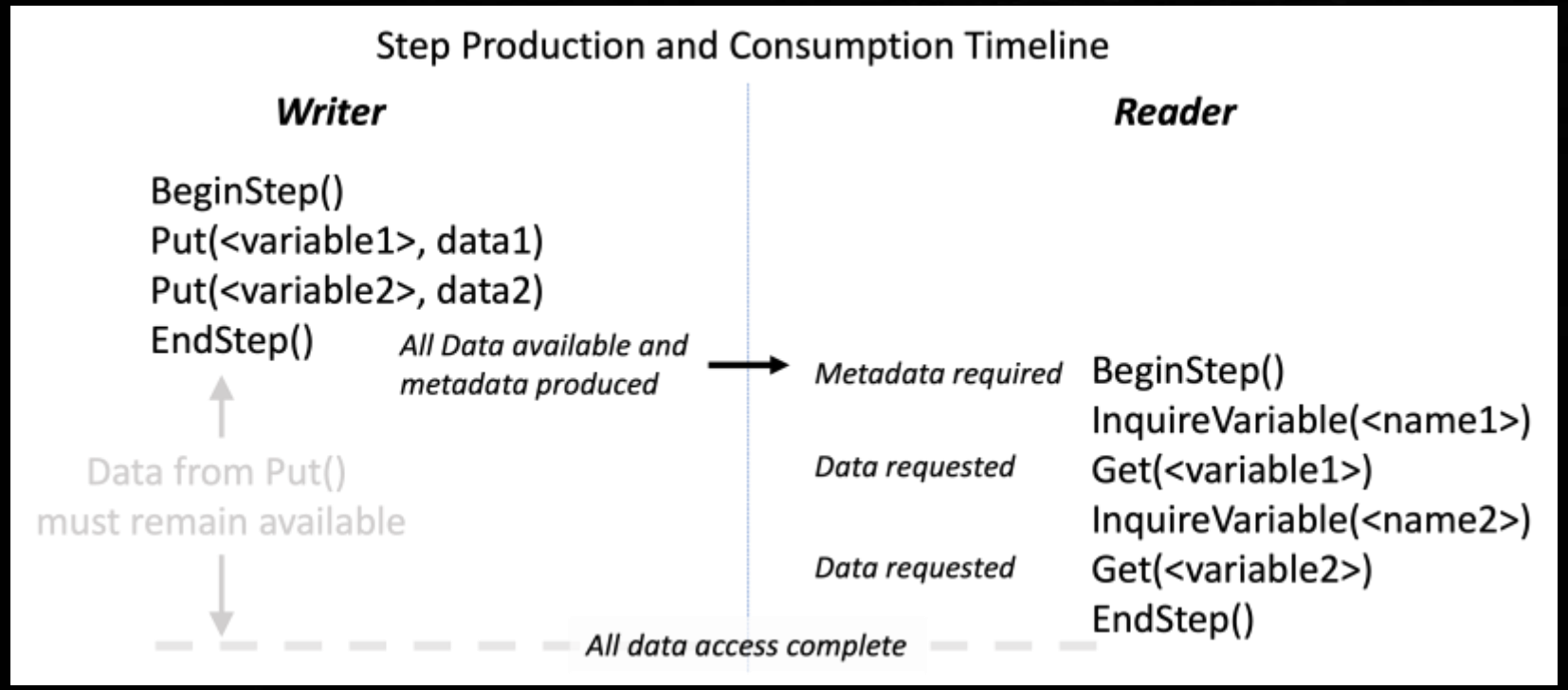
Requirements:

- Link writer and readers which are parallel (multi-core) programs
- Writer data geometry and reader data selections can change on each step
- Ensure that readers don’t impede writer progress or cause its failure
- Support dynamic connection and disconnection of reader applications
- Allow multiple simultaneous reader applications
- Move data at the highest speeds on HPC systems

pnorbert@ornl.gov

Challenges in coupling using a file-oriented API

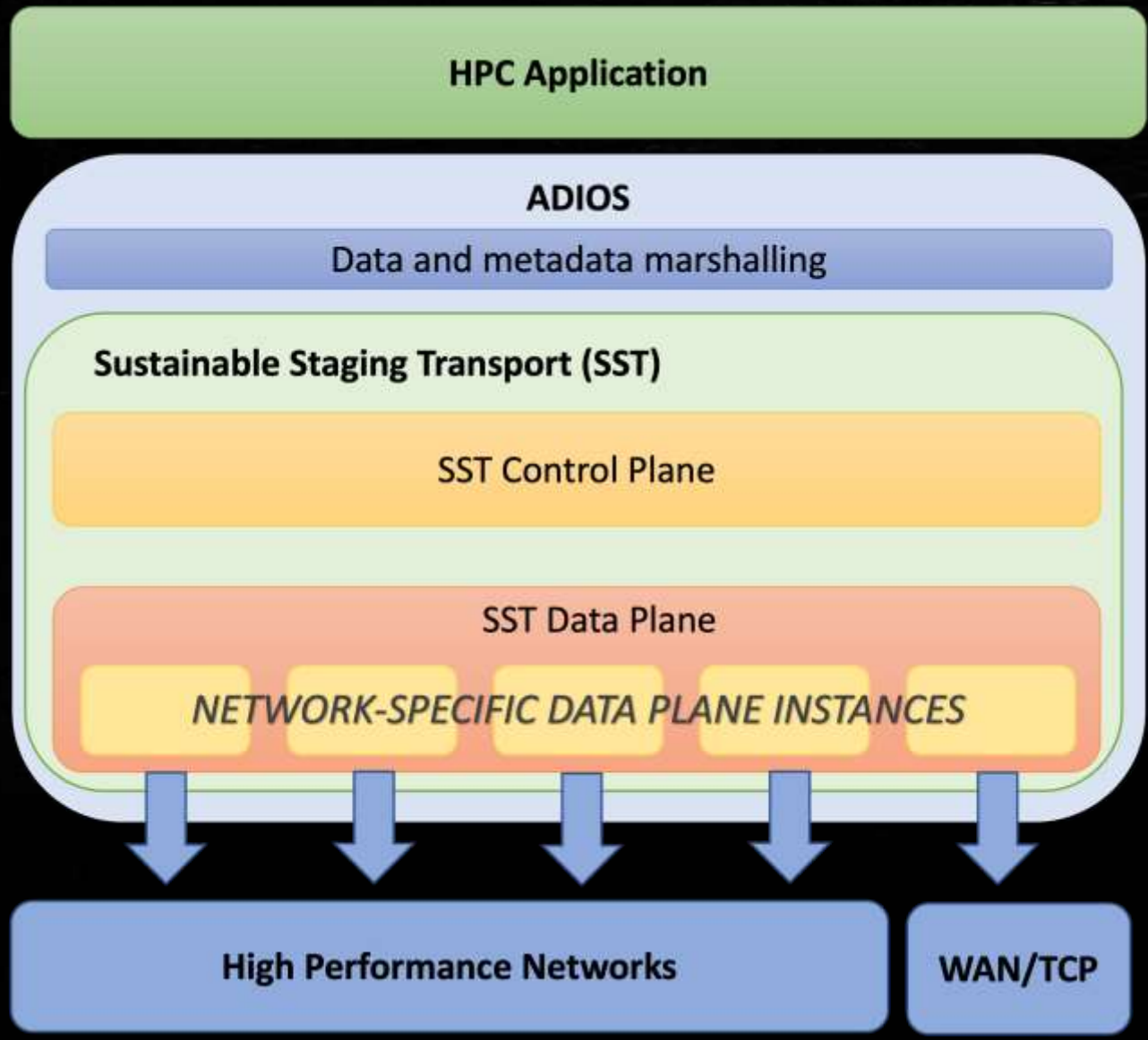
- ADIOS reader and writer APIs are highly flexible.
  - Writer can produce different data (variables, portions of global arrays, etc.) on every step
  - Reader can examine what is written (query metadata) before asking for specific data



  - Data queueing required in order to avoid delays on the writer side
  - Reader is largely asynchronous from the writer

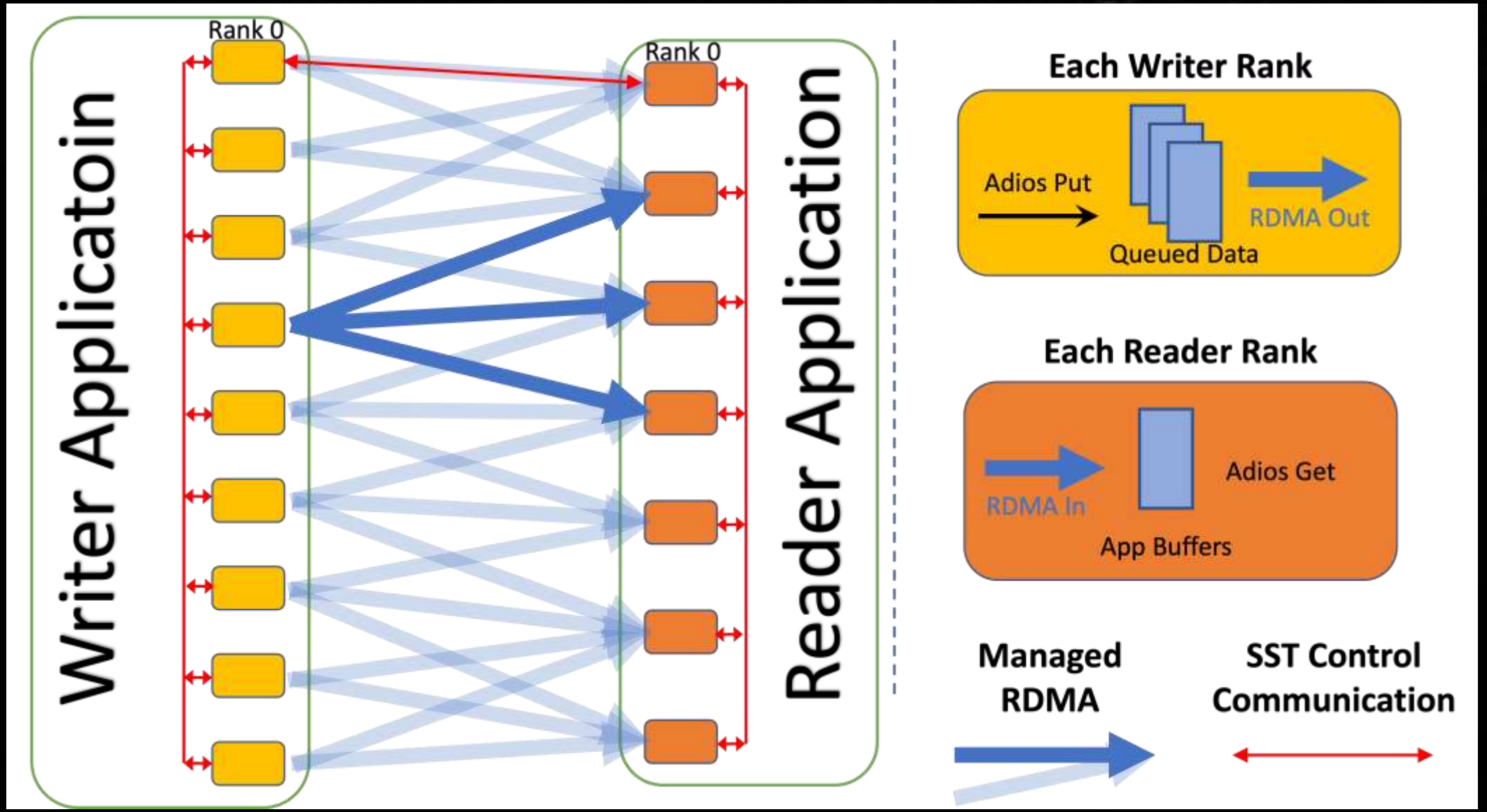
pnorbert@ornl.gov

SST Architecture



pnorbert@ornl.gov

SST Operation



pnorbert@ornl.gov

SST Control Plane

- High-level infrastructure responsible for startup, shutdown, and coordination
- Intra-application coordination done with MPI collectives
  - Collective operations occur in ADIOS Open(), Close(), BeginStep(), EndStep()
- Communication between writer and reader apps with message passing
  - EVPath IP-based messages between writer rank 0 and reader rank 0
  - Messages are asynchronous and read by a background network handler thread
- Coordinates with higher ADIOS layers to queue and release data and metadata
- Delegates all data operations to a network-specific data plane chosen at startup

pnorbert@ornl.gov

SST Data Planes

Custom interfaces to specific HPC network infrastructures

- TCP/IP (EVPath) data plane
  - WAN applications, application development
- Libfabric data plane
  - RDMA-capable transport where libfabric RDMA wrapper library is available
- UCX data plane
  - RDMA-capable where UCX is available
- MPI data plane
  - Based on MPI *inter-communicators*
  - Application must use thread-safe MPI
  - Limited to MPICH only
  - Sometimes better than Libfabric or UCX

pnorbert@ornl.gov

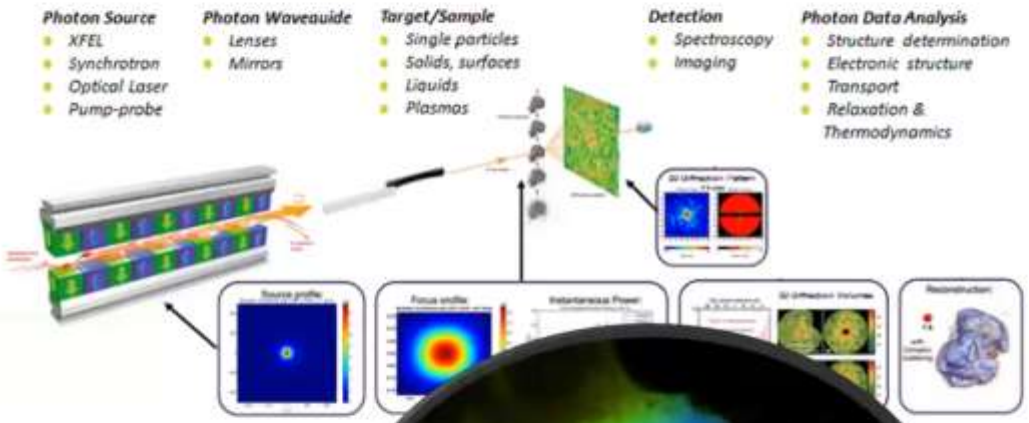
Performance Results

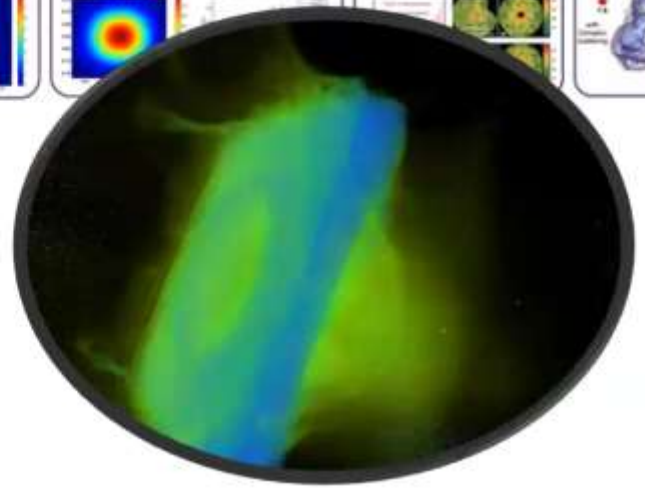
- Large: PIConGPU
  - Inter-application communication throughput at scale
- Medium: WDMApp
  - Percentage of I/O Overhead relative to main loop cost
- Small: WRF
  - Compare time to completion - ADIOS SST and PNetCDF

pnorbert@ornl.gov

In-memory coupling and online analysis on Top 10 HPC systems

Coupling several codes for a full digital twin of HIBEF experiments





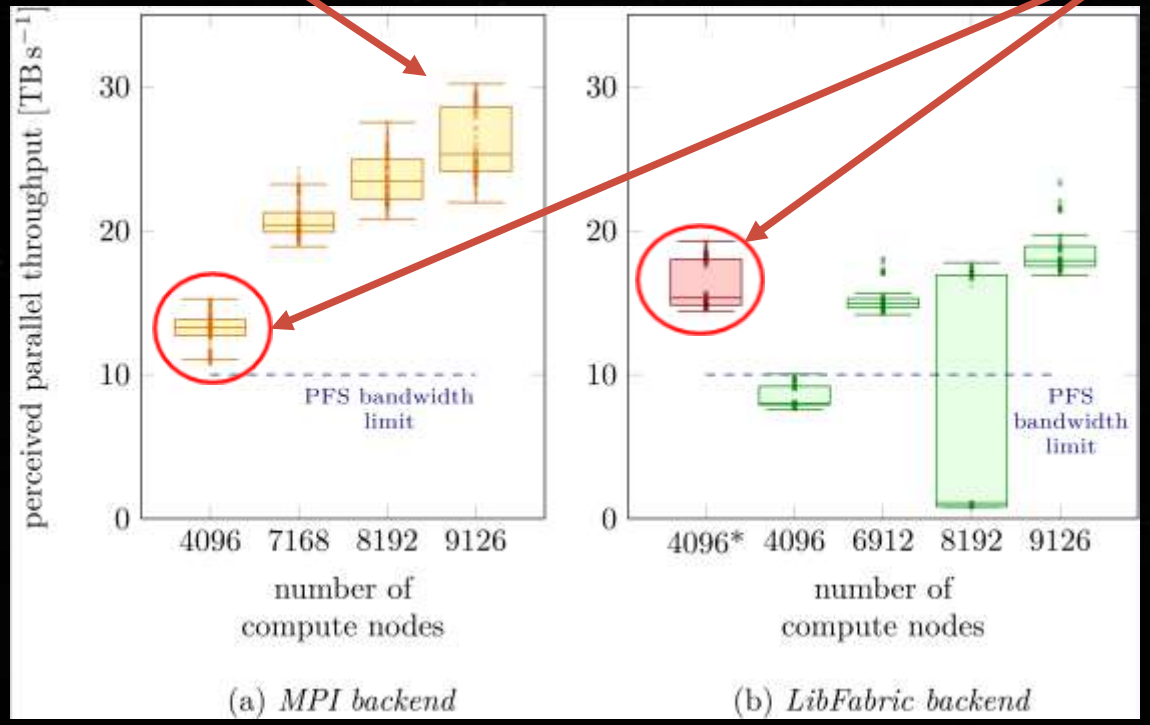
Run full digital twin of planned experiment on Top 10 HPC

- Use openPMD API with ADIOS2
- Add interactive, in-memory analysis
- Enable interactive simulation steering

PIConGPU: Kelvin-Helmholtz instability simulation on Frontier

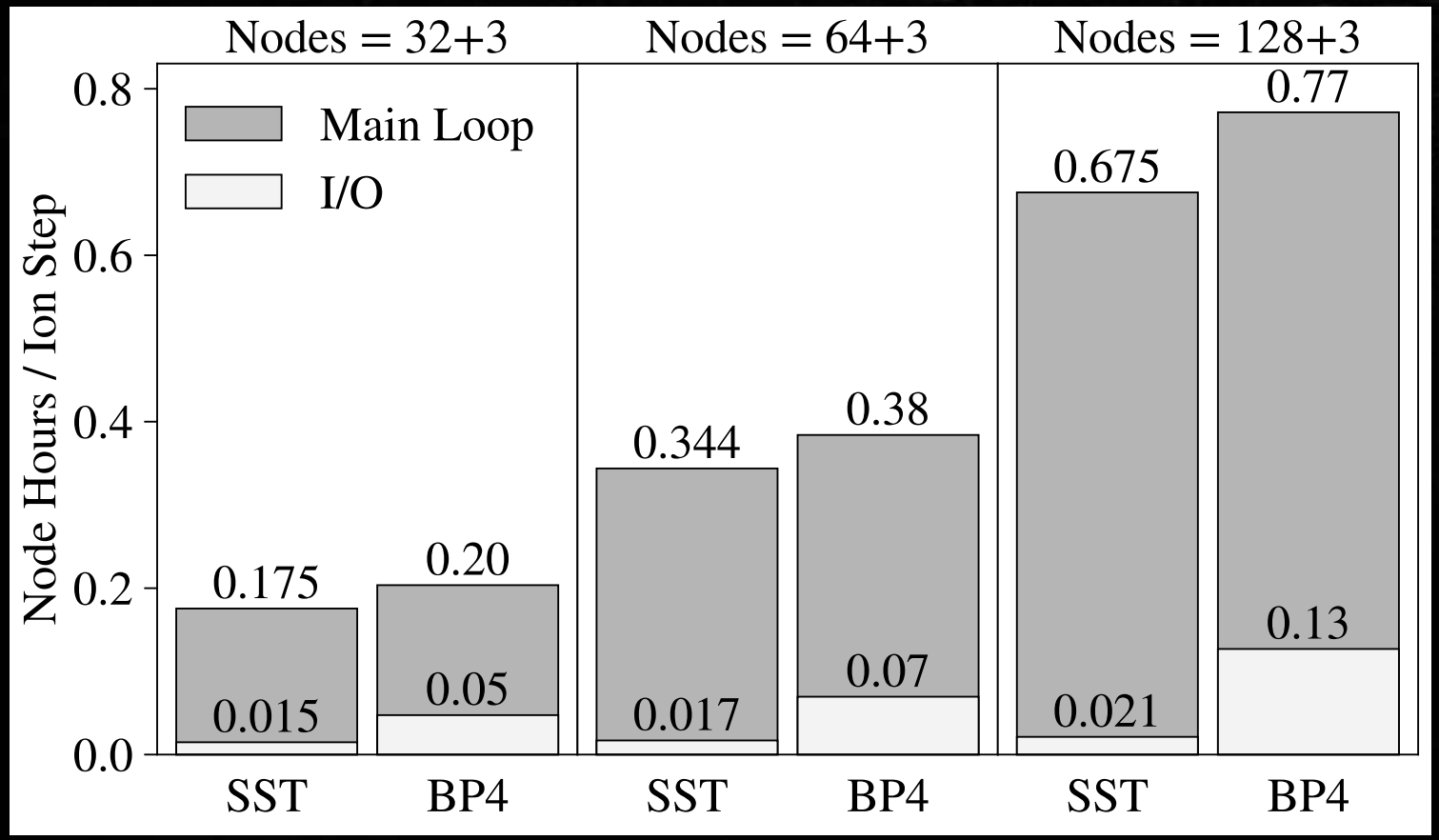
30 TiB/sec peak!

Promising at ½ scale



Peak filesystem performance is 10 TiB/sec – SST outperforms by 3x
Libfabric data plane needs tuning at higher scales

XGC+GENE I/O overhead vs compute on Crusher

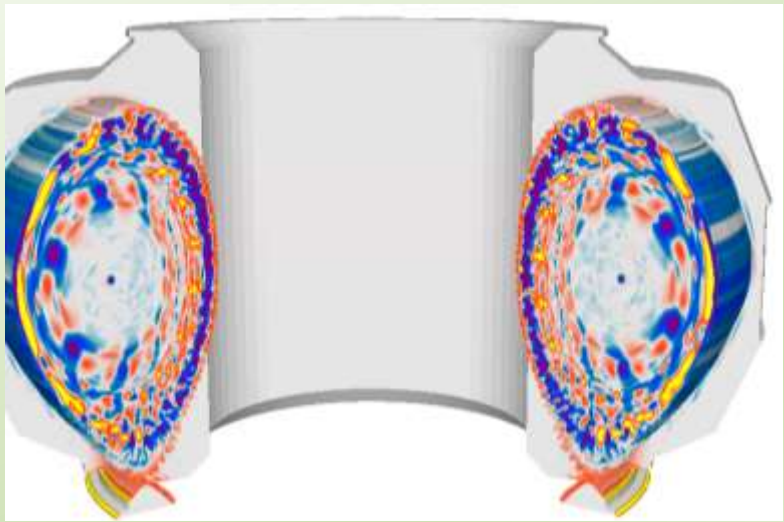


File-based BP4 coupling has considerably higher overhead than SST (MPI dataplane)

pnorbert@ornl.gov

HPC plasma physics workflows consist of large simulations, analysis, and visualization

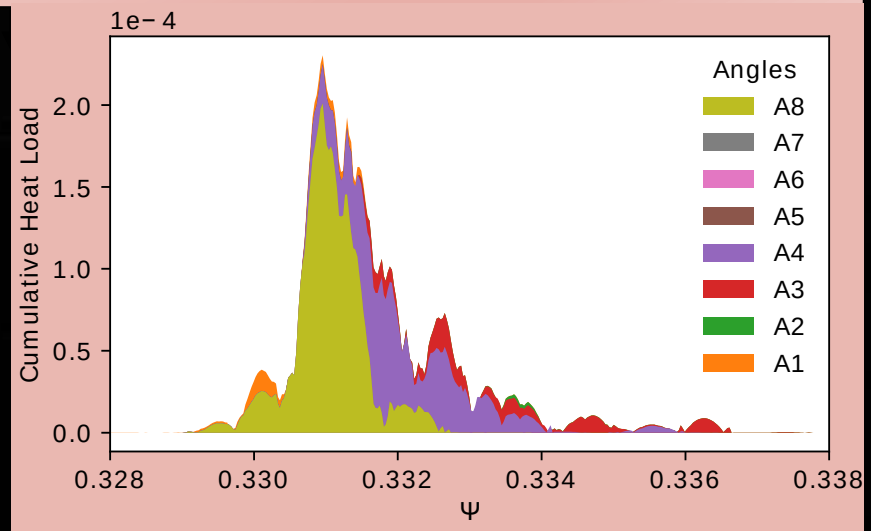
Advanced Simulations



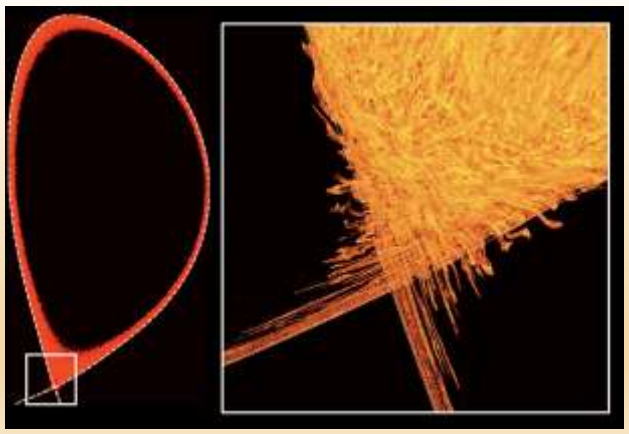
High Performance Analytics

Data: N tagged particles at step t
Result: Heat load calculation
Heatload $\leftarrow 0$
for $i \leftarrow 1$ *to* N **do**
  if P_i *is escaped* **then**
    UpdateParticleDB(P_i, t)
  end
  if P_i *is in divertor* **then**
    P_{esc}, $A_o \leftarrow$ SearchParticleDB(P_i, t)
    $\Gamma_{heat,A_o} \leftarrow$ HeatLoadCalc(P_{esc}, A_o)
  end
end
// Distribute and merge Particle DBs
AllGather(ParticleDB)

Algorithm 2: Heat load calculation algorithm.

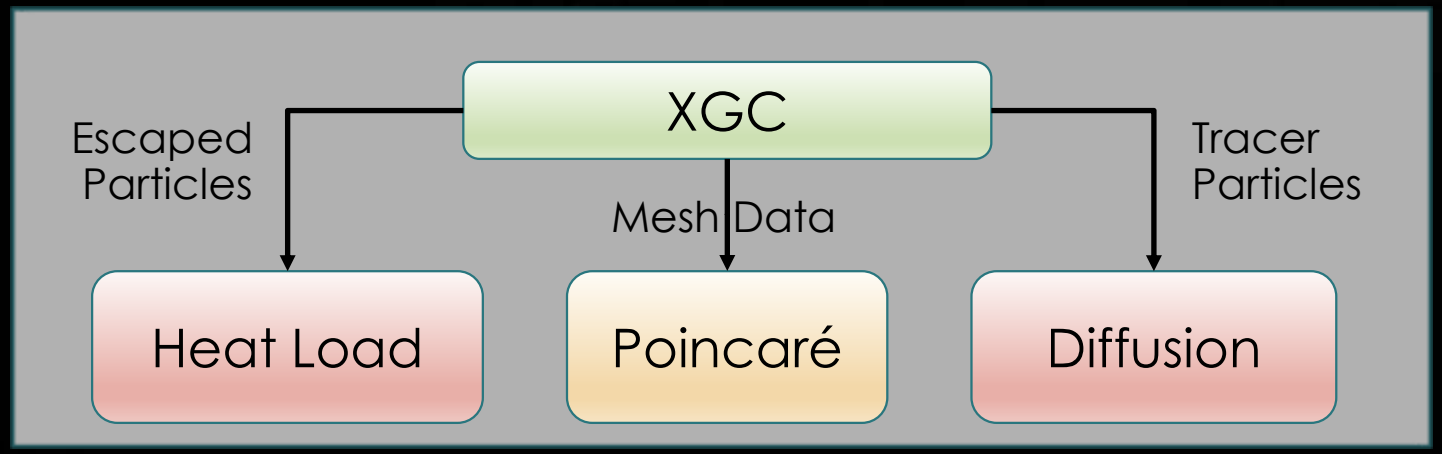


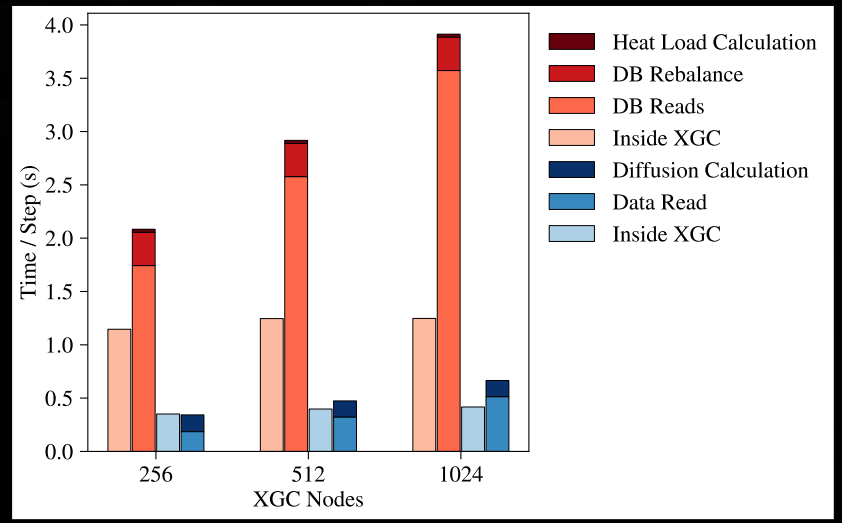
Visualization



Additional analyses are coupled to the simulations to extract new scientific insight

- In collaboration with XGC physicists, three were identified of particular interest
  - Poincaré puncture plot: Visualizes dynamic magnetic field structure
  - Heat load calculation: Measures energy deposition on the divertor plates
  - Diffusion calculation: Estimates particle transport properties
- These all inform better understanding magnetic confinement
- Run each as its own service, to offload as much as possible from XGC itself
  - Connect with ADIOS RDMA data movement

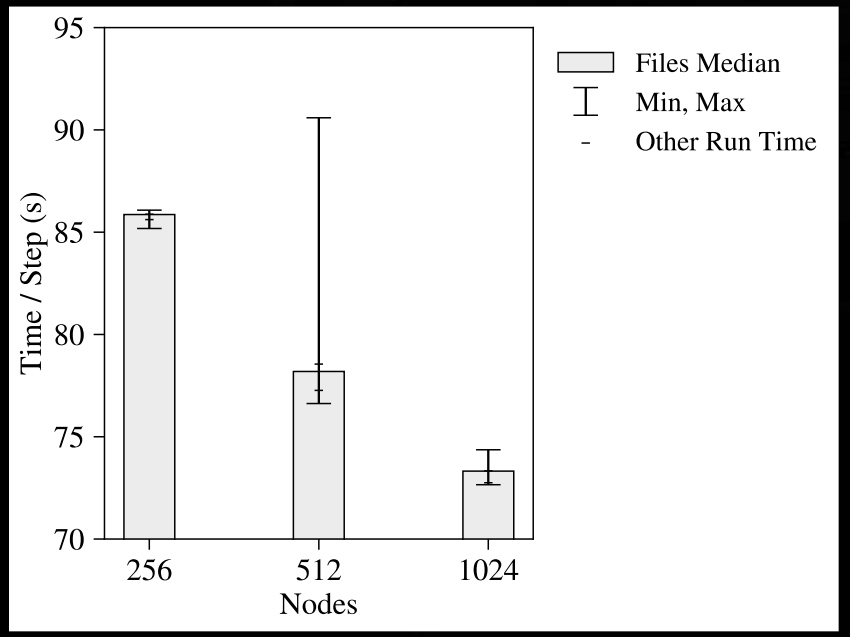


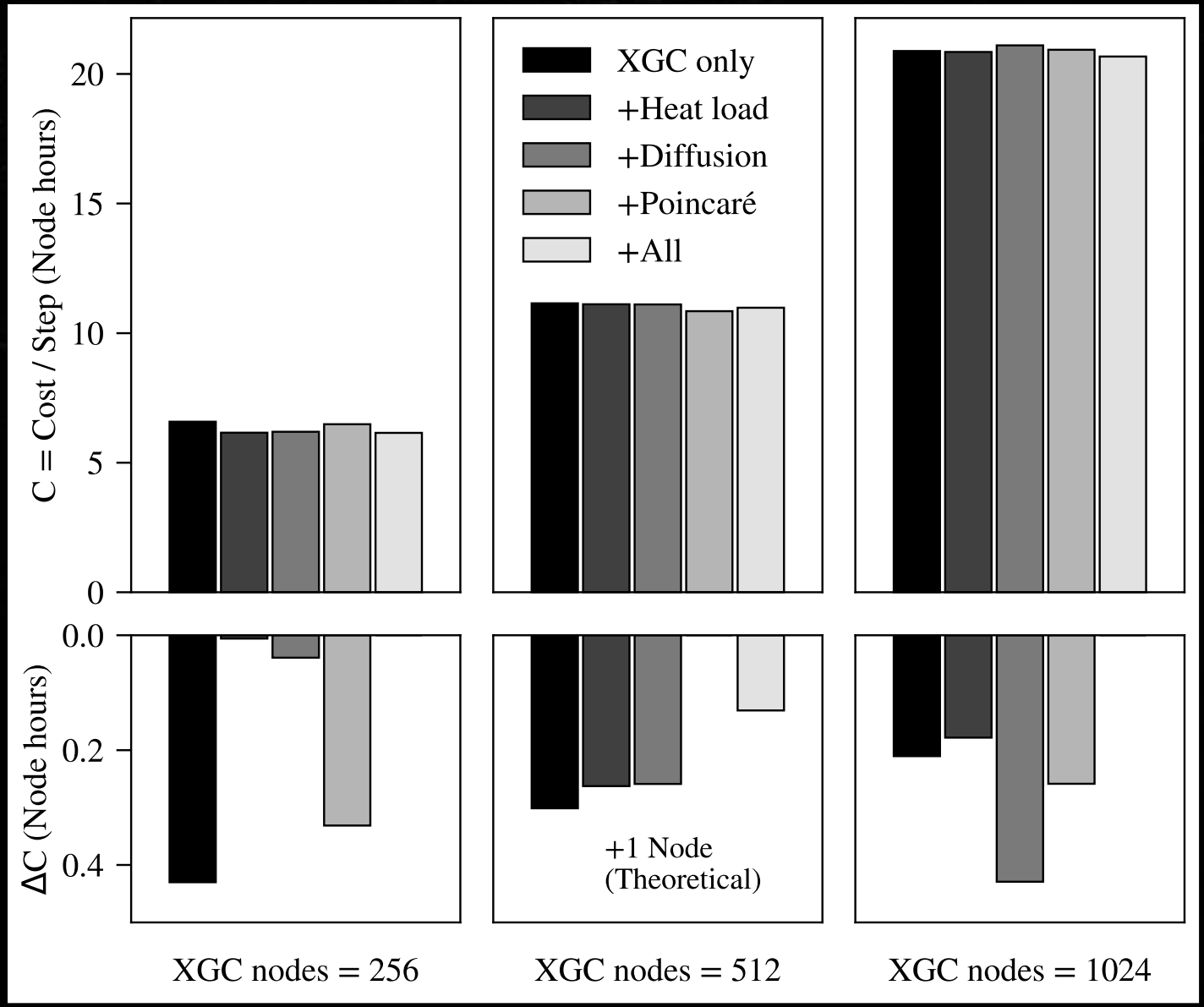


pnorbert@ornl.gov

Results: 0.1% cost to deliver three analysis in situ

- 1 extra node at all scales
- No significant impact to job cost to add the analysis
- Often below the variance of shared supercomputer

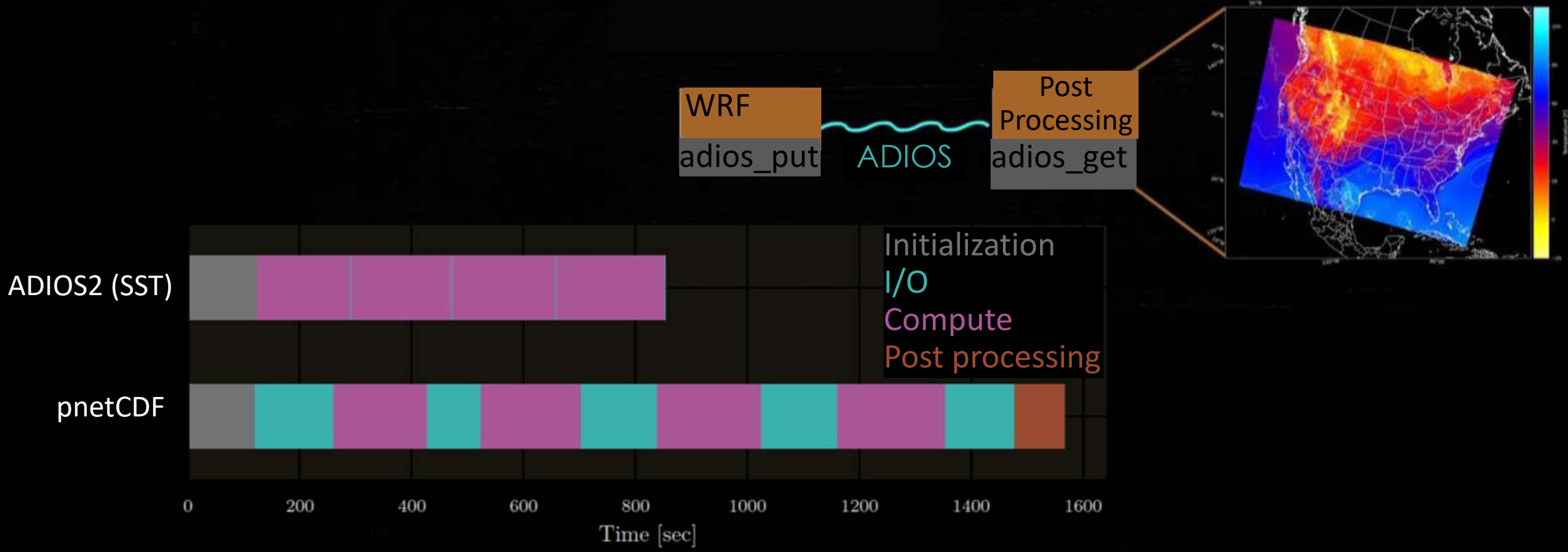




WRF-ADIOS In situ Analysis: M. Laufer, E. Fredj (2022)

Erick Fredj@2022
e-mail: fredj@jct.ac.il

- Post processing analysis pipeline built using ADIOS2 high-level Python API
- Data is processed as soon as it is available over the network (not filesystem)



Laufer, Michael, and Erick Fredj, "High Performance Parallel I/O and In-Situ Analysis in the WRF Model with ADIOS2." arXiv preprint arXiv:2201.08228 (2022).

pnorbert@ornl.gov

Questions