Sandia National Laboratories

Exceptional service in the national interest

AN HPC-ORIENTED RUNTIME ENVIRONMENT FOR ENABLING COMPUTATIONAL STORAGE

Matthew L. Curry

Computational I/O Stack Workshop
UC Santa Cruz
August 17, 2023

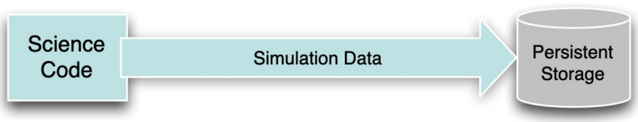
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

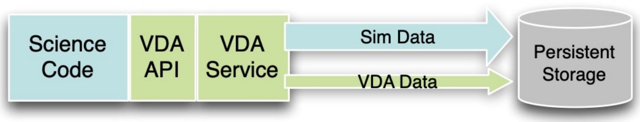
NNSA

IN TRANSIT WORKFLOWS: PROCESSING DATA OUTSIDE OF APPLICATIONS

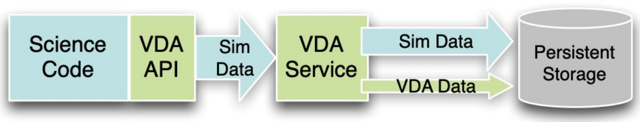
- "Extreme-Scale CoProcessing: An Evaluation of In Situ and In Transit Analysis"
  - Oldfield et al., 2013
- "This paper provides a comprehensive evaluation of in situ, in transit, and traditional post-processing workflows for an application that detects material fragments from data generated by a large-scale shock-physics simulation."
- Service-oriented in transit viz and data analysis (VDA) conforms with a computational storage vision
  - Reduced on-host data processing
  - Reduced data movement
  - Flexibility of runtime to make decisions



(a) Traditional post-processing VDA.



(b) Embedded in situ analysis VDA.



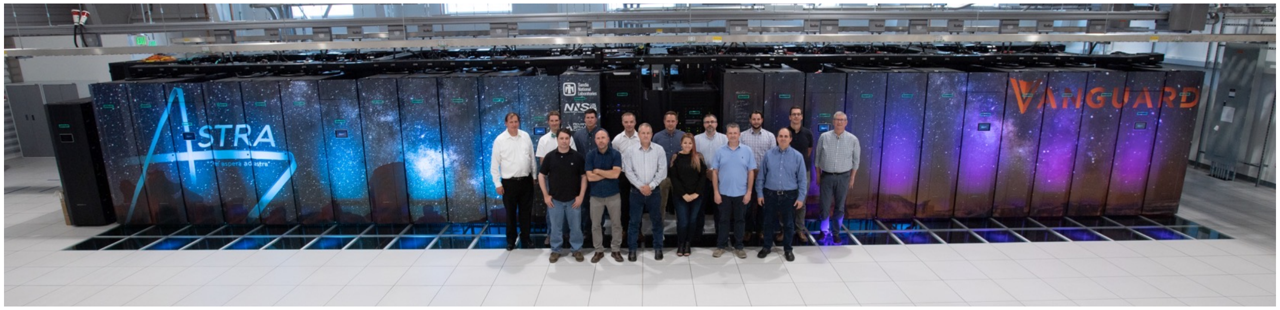
(c) Service-oriented in transit VDA.

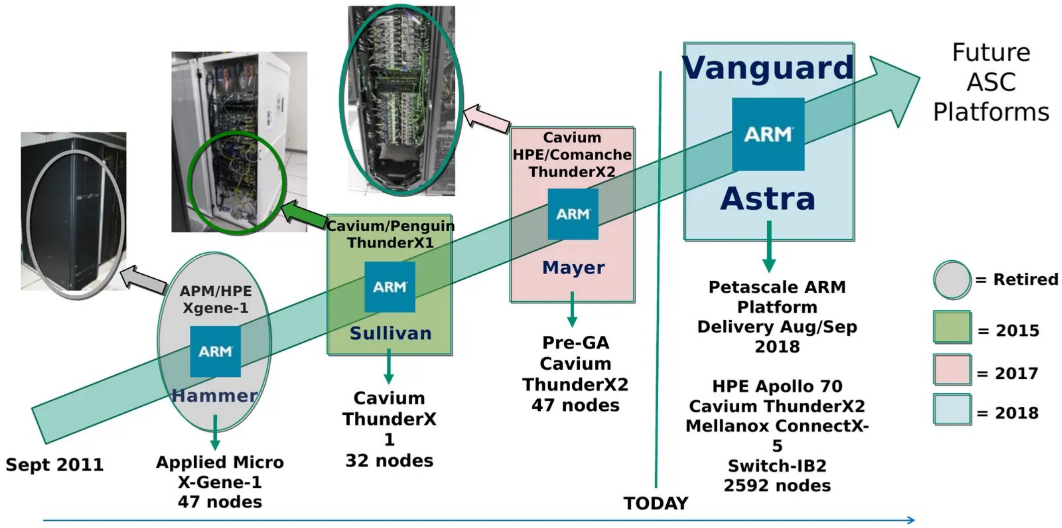
Fig. 1: Traditional and emerging workflow diagrams showing the flow of information from simulation to persistent storage.

PATHS TO ENABLING PROCESSING IN COMPUTATIONAL STORAGE

- Typical in transit analytics services will include a number of “interesting” components
  - vtk
  - I/O libraries (HDF5, CGNS, EXODUS, NetCDF, seacas, …)
  - Python
  - AI/ML frameworks
  - RDMA? MPI?!
  - Whole subsets of applications (inc. fftw, blas, C++20/F77, etc.)
- Very easy to find/provide on conventional HPC architectures (x86-64, NVIDIA)
- Computational storage devices don’t usually have these types of devices
  - Architectures not designed for code compatibility
  - Arm, RISC-V, MIPS (?!), all with different capabilities
- How might one port to such devices?

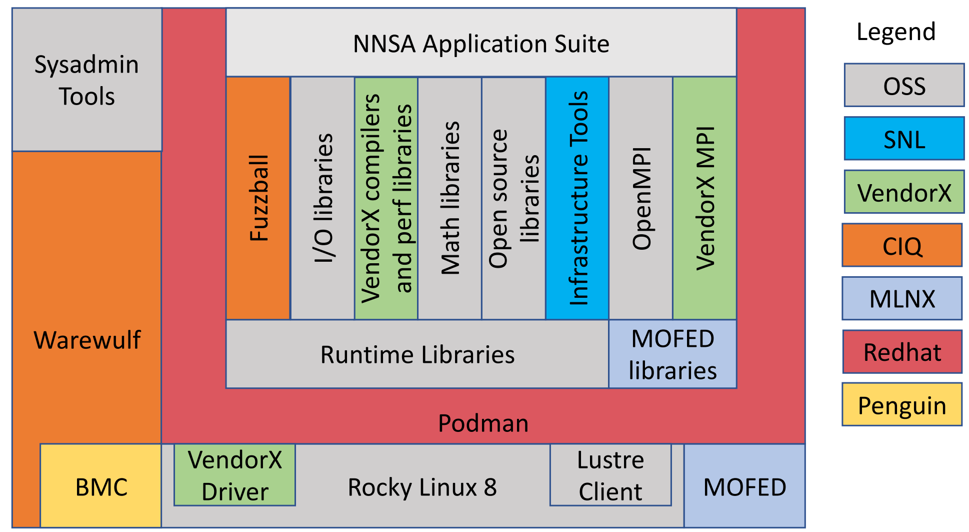
VANGUARD I/ATSE DEVELOPMENT HISTORY





ATSE: THE ADVANCED TRI-LAB SOFTWARE ENVIRONMENT

- Curated HPC software stack
  - Provides base set of compilers, MPI implementations, third-party libraries, tools, and other components known to work well together
  - Focused on needs of Sandia / DOE / NNSA / ASC codes
- Especially important for immature technologies
  - Many bugs, broken packages, and missing functionality
  - Need to do more to help users, avoid duplicated work
- Look and feel similar to OpenHPC, adapted for ASC:
  - Add missing packages (e.g., ParMETIS, CGNS)
  - Add microarchitecture and compiler optimizations
  - Port to Spack for dynamic builds on rare hardware



ATSE Provides "Ready to Go" Programming Environment for ASC Codes on Novel Hardware

ATSE: AN EMINENTLY PORTABLE STACK

- ATSE has been ported to a wide variety of niche architectures, including:
  - Several Arm variants (Preproduction chips, ThunderX2, A64FX, Neoverse, **Bluefield 2**, etc.)
  - Dataflow and traditional accelerators
  - RISC-V (both real chips and FPGA-resident)
- ATSE embodies a framework and set of specs that allow quick portability
  - Two person-days to two person-weeks for initial port to a new architecture
- Batteries included for a small set of demonstrator applications
  - Manageable set of packages
  - High impact
- Container and native installation paths
- Easy to iterate for application needs

THANKS