PandAna: An environment for scalable high-level HEP analysis on HPC

Achievements

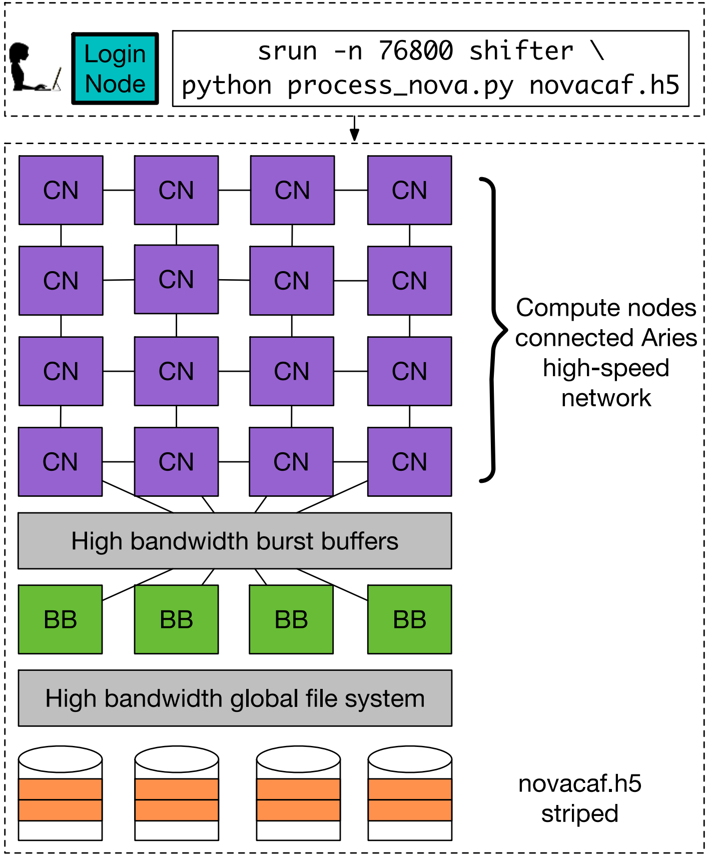
Demonstration of scalable parallelization of an analysis code from NOvA by replacing serial IO mechanism with parallel IO.

Significance and Impact

Allows existing analysis code developed by experimenters to be deployed at HPC sites for processing of large datasets.

Research Details

- Provide an easy-to-use environment for fast and scalable HEP high-level data analysis
  - Users can develop on laptops or local clusters and deploy code to HPC
- Use HDF5 for fast parallel reading of large amounts of data
- Use Python and popular Python data science tools (numpy, pandas)
- Introducing to HEP the “tidy data” analysis model, using large data matrices and distributed data parallelism
  - Use MPI for distributed parallelism
  - The parallelism in user code is implicit



```
srun -n 76800 shifter \
python process_nova.py novacaf.h5
```

Login Node
CN
Compute nodes connected Aries high-speed network
High bandwidth burst buffers
BB
High bandwidth global file system
novacaf.h5 striped