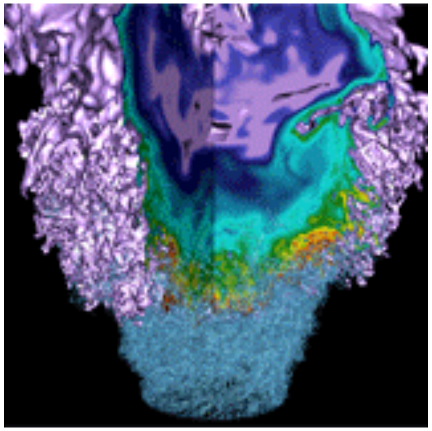
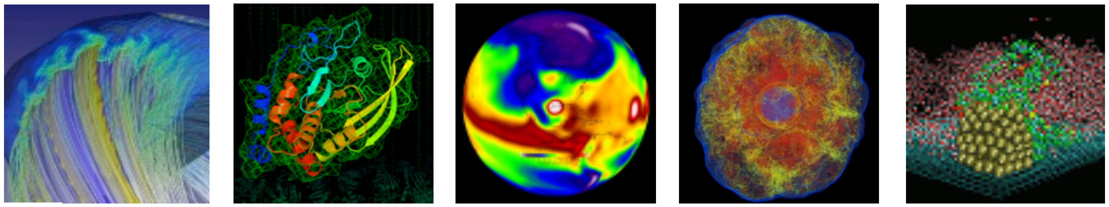
Cray DataWarp Administration & SLURM Integration





NERSC 40 YEARS at the FOREFRONT

Tina Declerck, NERSC
Iwona Sakrejda, NERSC
Dave Henseler, Cray

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Background

- Computing is a balancing act
- CPU, memory capacity, memory bandwidth, IO, network bandwidth, network latency
- Things are getting out of balance

Cray DataWarp Intro

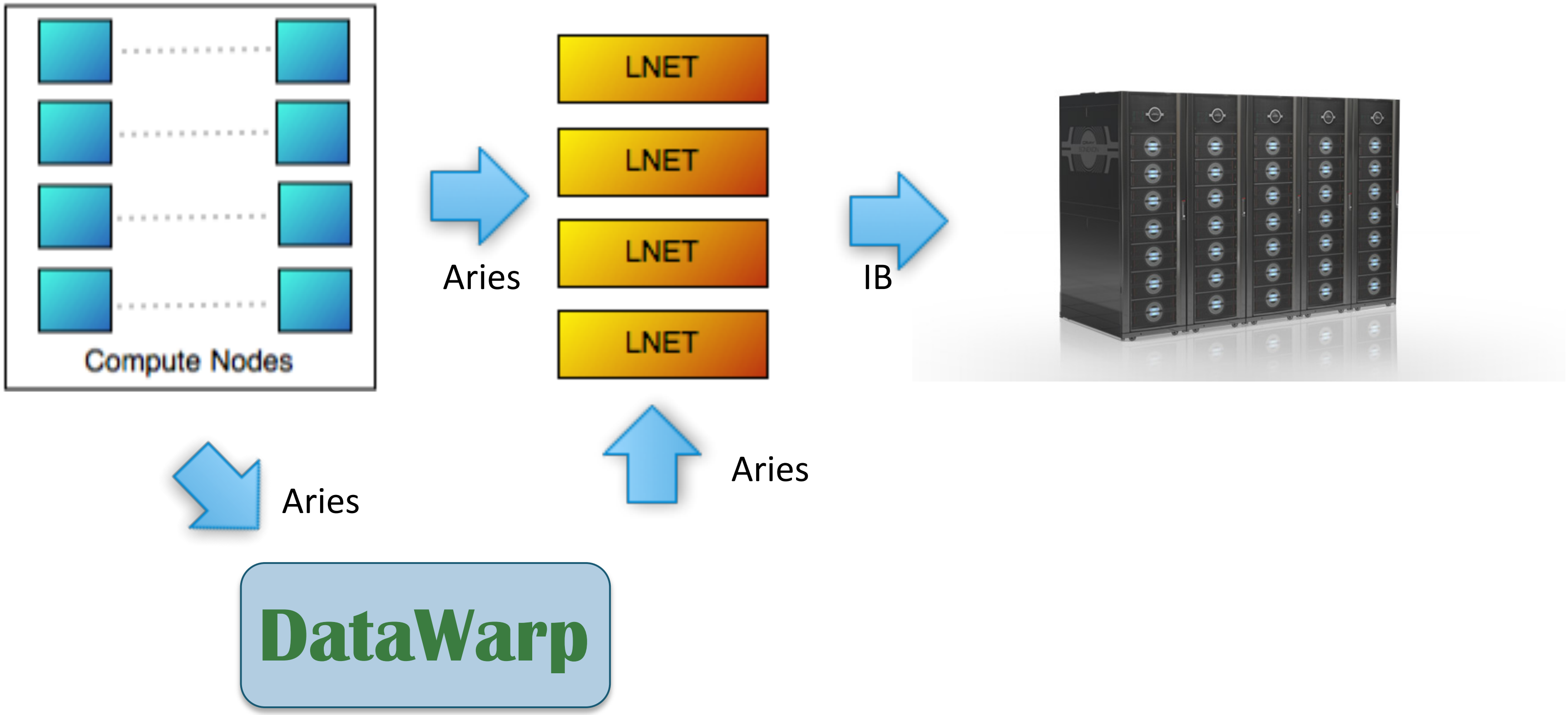
- **What is it?**

+

+ Software Development

Why use it?

- **Checkpoint / Restart**



Compute Nodes

Aries

LNET

LNET

LNET

LNET

IB

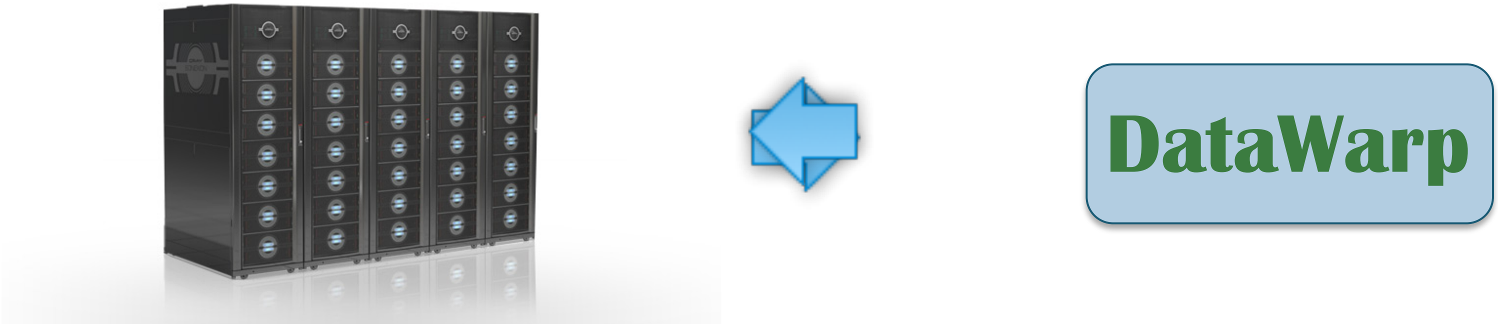
Aries

Aries

DataWarp

Use Cases

- **Pre and Post staging of files**



Use Cases

- **Other use cases**
  - Compound jobs
    - Multiple jobs or users access the same data
  - Implicit cache
    - Intermediary storage between RAM and disk
  - Private cache used as swap
  - Private scratch used as a /tmp
  - Stripe across multiple DW nodes
    - Additional space
    - Improved performance

DataWarp Phases

- **Phase 0**
  - Statically configured as swap or scratch
- **Phase 1**
  - dynamic allocation and configuration of DataWarp storage to jobs
  - job/application controlled explicit movement of data between DataWarp and PFS storage
- **Phase 2**
  - movement of data between DataWarp and PFS storage
- **Phase 3**
  - Ability to run applications on DataWarp server nodes

SSD Considerations

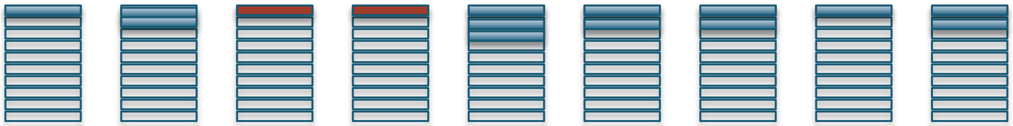
- **Consumable resource**
  - Based on drive (or diskful) writes per day (DWPD) for some specified time frame (usually 5 years).
  - Example:
    - Device size is 400GB
    - Listed as 3 DWPD
    - 5 year life

    DWPD * Device size * life in years * days per year = Data written

    3 * 400 * 5 * 365 = 2,190,000 GB can be written to the device
  - Can wear out a device in a relatively short period of time

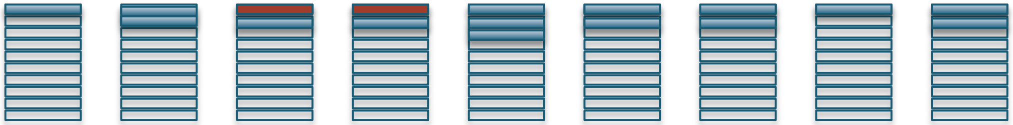
SSD Considerations

- **Wear leveling**
  - Balances block usage to ensure even use on an SSD
    - Dynamic - ensures new writes or re-writes are written to new areas on the SSD



SSD Considerations

- **Wear leveling**
  - Static – same as dynamic + relocates static files occasionally to free those blocks for additional writes



TRIM command

- **Required on both SSD and at OS level**
  - Identifies blocks that can be removed.
  - SSD can't over-write like a disk
  - Data written in pages but on SSD must be erased in blocks
    - Active data must be written to a different block so the block can be erased.
  - If not used it can affect performance over time

Admin for DataWarp nodes

- **In general hardware similar to other hardware**
  - No special monitoring needed at node level
- **Need to monitor**
  - Available life
  - Excessive use
  - Bit error rates
- **Query firmware / software levels**
- **Event logging**
- **Command line and API**

DataWarp Operating Modes

- **Understand access request types**
  - Job instance
  - Persistent instance
- **Two types of use for DataWarp**
  - Scratch
  - Cache
- **Access in three ways**
  - Striped
  - Private
  - Load balanced

WLM Admin Requirements

Why isn't my job running?

- **Query a DataWarp instance – job or persistent**
  - Owner, size, duration, parameters, owning job (if applicable), DataWarp nodes
- **Diagnostic information**
  - Same information as a query but based on a job id
  - Provide status of DataWarp nodes
- **Restrict access**
  - Limit number or space used by a single user
  - Limit access by a specific user or group
  - Limit access to only a specific user/group etc.

WLM Requirements

Other User Related Items

- **Modify existing DataWarp instance**
  - Duration
  - Size (if possible)
  - Add/Modify user access
  - Other parameters (TBD)
- **Provide DataWarp statistics for each job**
  - Bytes in/out per server

WLM Requirements
Maintenance Tasks

- **Kill an existing job or persistent DataWarp instance**
  - Why?
    - Node is in a bad state and needs to be fixed
    - TheDW instance is no longer in use but is still held
  - Data Considerations
    - Purge
    - Migrate
    - Drain
- **Disallow access – system maintenance**

Other

- **Attempt to wear level across all DataWarp nodes.**
  - Keep one from wearing out before the others

SLURM Specific Requirements

- **Job Prioritization**
  - DataWarp use should be included in the calculation for job prioritization
- **Advanced Reservations**
  - Allow DataWarp instances to be created in advance and requested by jobs as needed.
- **Resource Limits**
  - Allow to set resource limits on a per job, per user,

Conclusion

- **Cray's DataWarp will be a useful tool**
- **We want to ensure we have the data we need to provide support**
  - At the hardware level
  - At the job level

Acknowledgments

- **Iwona Sakrejda - NERSC**
- **Dave Henseler – Cray**

This work was supported by the Director, Office of Science, Office of Advance Scientific Computing Research of the U.S. Department of Energy under contract No. DEAC02-05CH11231

Thank you!

Questions?