als Baustein moderner Speicherhierarchien
Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

Copyright 2016 Cray Inc.
Trends in the Memory / Storage Subsystem

**Today**
- CPU
- Memory (DRAM)
- Storage (HDD)
- Distant Storage (WAN/Tape)

**Near Future**
- CPU
- Near Memory (HBM/HMC)
- Main Memory (DRAM)
- Far Memory (NVDIMM)
- Network NV Mem (SSD)
- MidStorage (HDD)
- Distant Storage (Object/WAN/Tape)

- On Node
  - 100+ µs Flash
  - \( O(1\mu s) \) NVRAM

- Off Node
Overview - What is DataWarp?

- DataWarp is Cray’s implementation of the Burst Buffer concept, plus more
- Has both Hardware & Software components
- Hardware:
  - XC40 Service node, directly connected to Aries network
  - PCIe SSD Cards installed on the node
- Software:
  - DataWarp service daemons
  - DataWarp Fileysystem (using DVS, LVM, XFS)
  - Integration with WorkLoad Managers (Slurm, M/T, PBSpro)
Cray XC System Environment

Cray XC Supercomputer

Boot RAID

SMW

StorageSwitch Fabric

MDS

Lustre OSS

Lustre OSTs – global work

Management Server

Visualization Server

Pre- & Post-processing

NAS - home

Login Servers

Login Servers

IB Fabric

Data Mover

Cray XC System Environment

Lustre OSTs – global work

NAS - home

Login Servers

Management Server

Visualization Server

Pre- & Post-processing

Cray Proprietary

Compute nodes

MOM Nodes (SIO)

Network Nodes (SIO)

LNET Router Nodes for Lustre (SIO)

DVS Server Nodes for NGF (SIO)

Boot, Syslog and System Database Nodes (SIO)
Cray XC System Environment

Cray XC Supercomputer

Boot RAID

SMW

StorageSwitch Fabric

Login Servers

Login Servers

Lustre OSS

Lustre OSTs – global work

MDS

Management Server

Visualization Server

Pre- & Post-processing

IB Fabric

Data Mover

Cray XC System Environment

DataWarp nodes
Compute nodes
MOM Nodes (SIO)
Network Nodes (SIO)
LNET Router Nodes for Lustre (SIO)
DVS Server Nodes for NGF (SIO)
Boot, Syslog and System Database Nodes (SIO)
DataWarp Hardware Setup

2 nodes per blade and 2 SSDs per node

$ xtnodestat
C0-0
n3 ----
n2 SSSSSSS----
n1 SSSSSSS----
c0n0 ----
s0123456789abcdef
Use Case: Local Storage on Demand

Per Node Scratch

- Each compute node in a job is assigned a private part of the allocated SSD space
- Much faster than “faking it” with a parallel file system

Per Node Swap Space

- Dynamic compute node swap space
Use Case: Shared Fast / SSD

Shared Fast Scratch

• High Bandwidth access to shared files
• Files can be striped across multiple DataWarp Nodes
• Space can be temporary for the job, or be marked as persistent to work between jobs
Use Case: Checkpoint / Restart

Fast Checkpoint / Restart

- User asks for enough SSD to cover the number of concurrently resident checkpoints
- High Bandwidth checkpoints are written to SSDs
- Followed by an asynchronous explicit or transparent copy out to rotating storage
Use Case: File System Caching

Transparent File System Caching

- Global file system caching
- Both on-demand and transparent to the application
- Phase 2 Feature
DataWarp – Minimize Compute Residence Time

Key:
- Compute Nodes
- Compute Nodes - Idle
- I/O Time Lustre
- I/O Time DW
- DW Nodes

Time (Lustre Only):
- Initial Data Load
- Compute
- Final Data Writes

Time (DataWarp):
- DW Preload
- DW Post Dump

Node Count (Vertical):
- Initial Data Load
- Compute
- Final Data Writes

Node Count (Vertical):
- DW Preload
- DW Post Dump

Timestep Writes

Timestep Writes (DW)
Slurm Job Script Commands Simple Example: With and Without DataWarp

#!/bin/ksh
#SBATCH -n 3200 -t 2000
export TMPDIR=/lustre/my_dir
srun -n 3200 a.out

#DW jobdw type=scratch access_mode=striped capacity=1TiB
#DW stage_in type=directory source=/lustre/my_dir destination=$DW_JOB_STRIPED
#DW stage_out type=directory destination=/lustre/my_dir source=$DW_JOB_STRIPED
export TMPDIR=$DW_JOB_STRIPED
srun -n 3200 a.out
12 Million Random 4K IOPS!

140 DataWarp Nodes
4k random writes and reads
4480 1GiB Files
World Record IOR Result – KAUST with DataWarp

Data Warp Performance

- 264 DataWarp Nodes
- 4000 Compute Nodes
- Shared Scratch IOR Test
- 1.5 TB/secWrites
- 1.8 TB/sec Reads
DataWarp Documentation

- DataWarp Installation and Configuration Guide S-2547-5204
  - This publication covers the installation procedure for DataWarp SSD cards as well as post-boot configuration; it is intended for system administrators.

- DataWarp Administration Guide S-2557-5204
  - This publication covers administrative tasks for Cray XC™ series systems installed with DataWarp SSD cards; it is intended for system administrators.

- DataWarp User Guide S-2558-5204
  - This publication covers DataWarp commands, DataWarp job script commands, and the DataWarp API and is intended for users of Cray XC™ series systems with DataWarp SSD cards.
When you need to know more than just the temperature.

Cray Inc.

www.cray.com