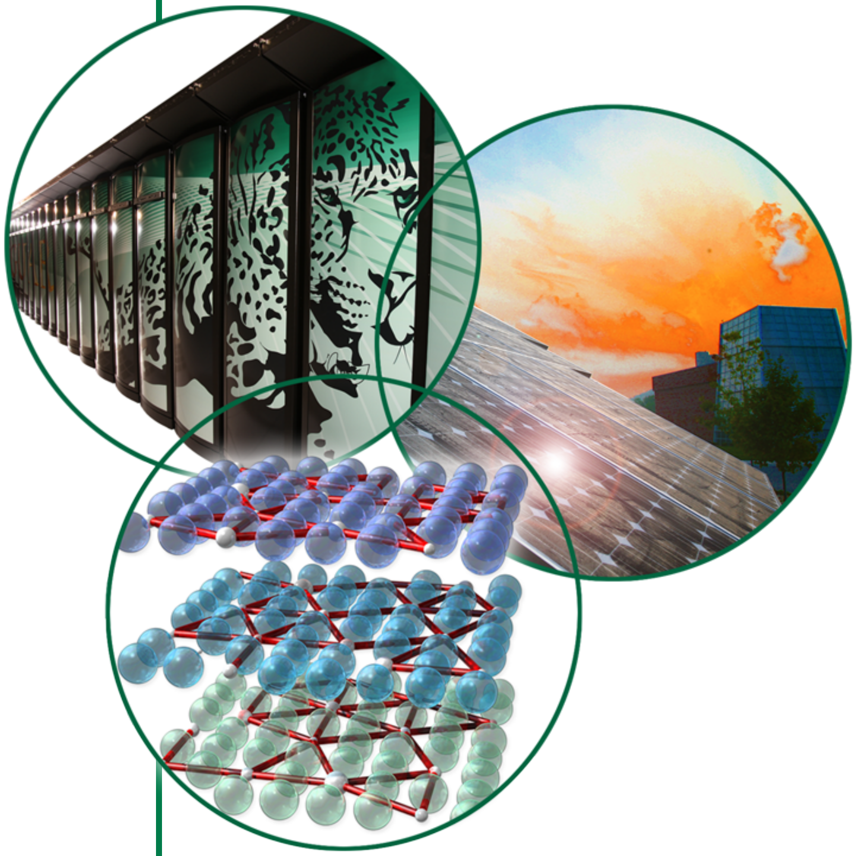
# Architecture and Implementation of Lustre at the National Climate Computing Research Center



**Douglas Fuller**

National Climate Computing Research Center / ORNL

LUG 2011

U.S. DEPARTMENT OF ENERGY

NCRC

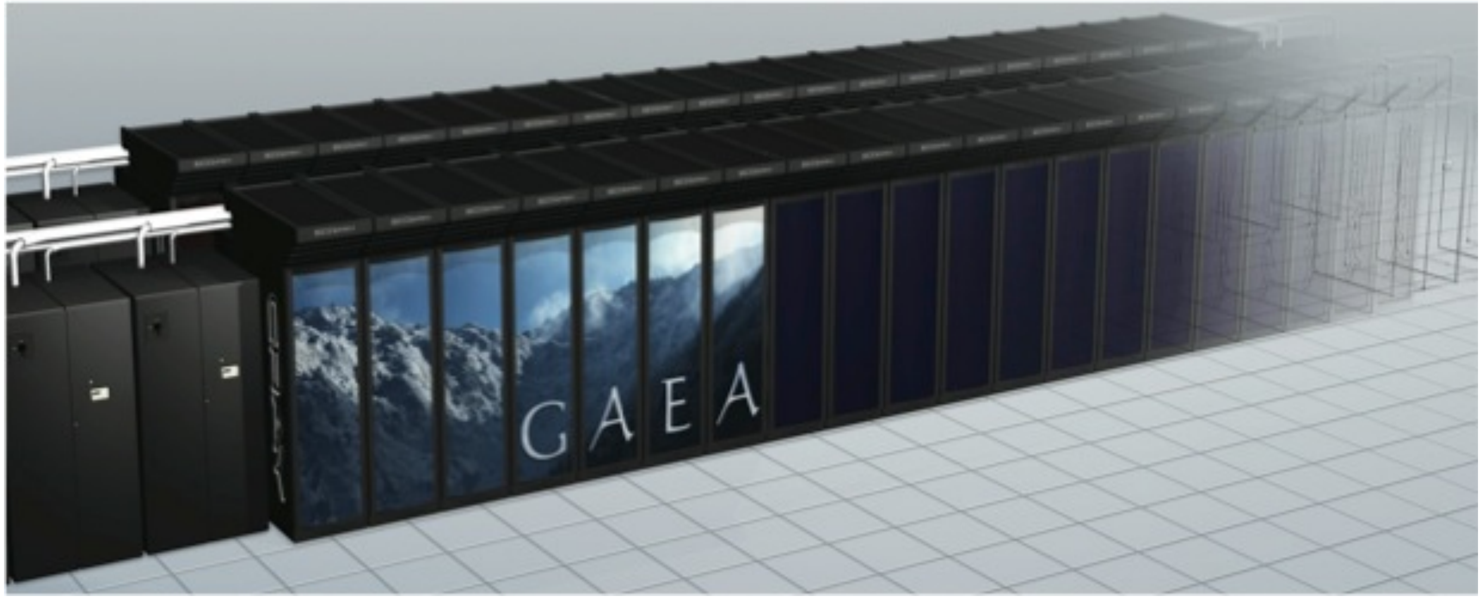
OAK RIDGE National Laboratory

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

# About NCRC

- Partnership between Oak Ridge National Laboratory (USDOE) and NOAA (USDOC)
- Primary compute and working storage (Lustre!) at ORNL (Oak Ridge, TN)
- Primary users at Geophysical Fluid Dynamics Laboratory (Princeton, NJ)
- Operational in September 2010
- Upgrades through 2012, scheduled operations through 2014

# Gaea Compute Platforms



- Phase 1: Cray XT6
  - 2,576 AMD Opteron 6174 (“Magny-Cours”) processors
  - 260 TFLOPS
  - 80 TB main memory
  - Upgrade to XE6/360 TF in 2011

- Phase 2: Cray XE6
  - 5,200 AMD Opteron 16-core (“Interlagos”) processors
  - 750 TFLOPS
  - 160TB main memory

# File System Design: Requirements

- The obvious (capability, capacity, consistency, cost)
- Consistent performance
  - More production-oriented workload
  - High noise from compute and auxiliary functions
- Resiliency
  - Local component failures (nothing new)
  - Compute partition failures
  - WAN connectivity issues

# System Specification

- Capability: projections from previous systems
  - Aggregate daily data production
  - Current code I/O duty cycles
  - Overhead from auxiliary operations
  - Delivered I/O from two primary partitions
- Capacity: fit the use cases that need performance
  - Scratch
  - Hot dataset cache
  - Semi-persistent library
  - Staging and buffering for WAN transfer

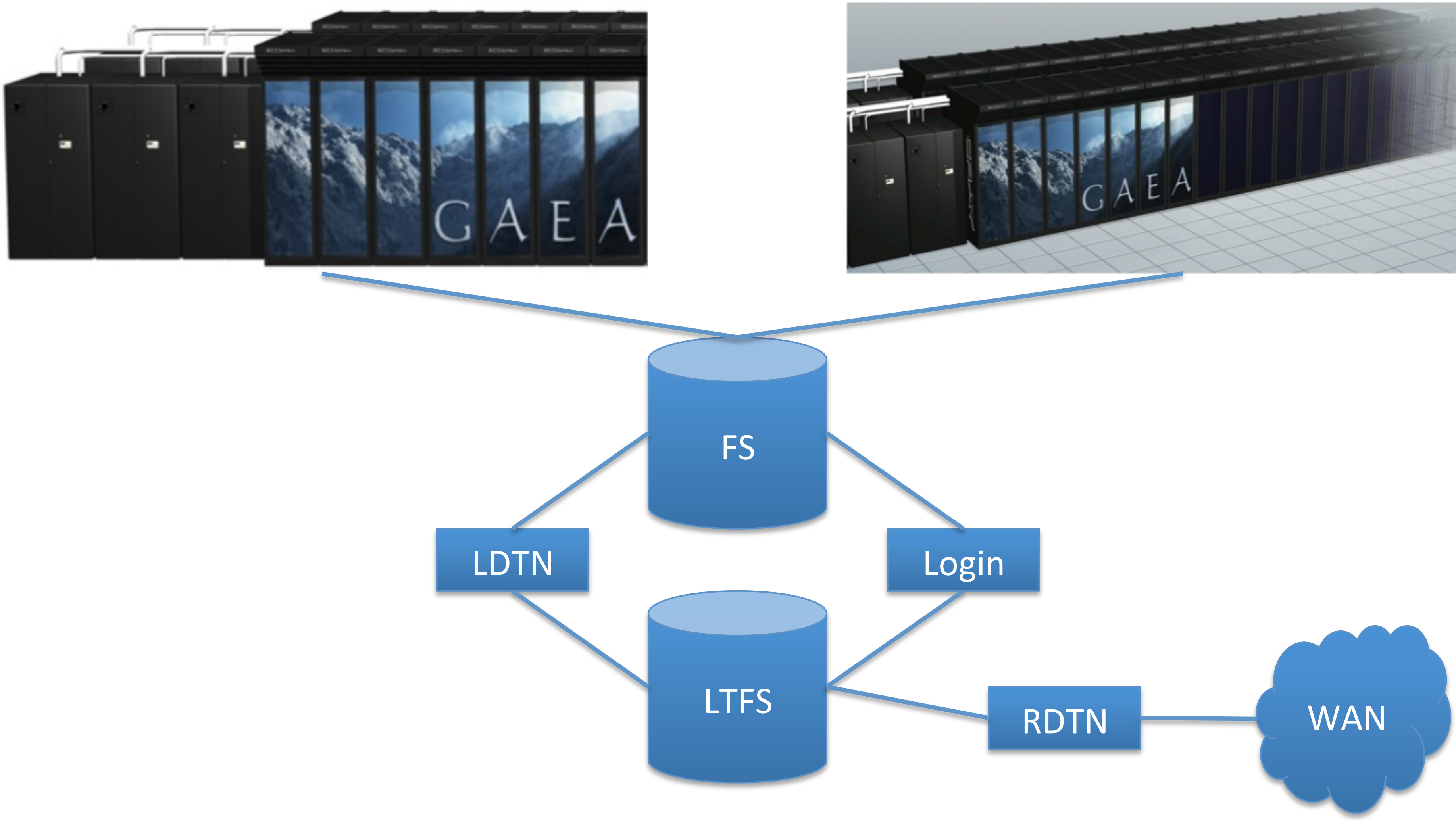
# System Specification

- Consistency: use cases increase variability
  - Some demand capability (scratch, hot cache)
    - Significantly more random access
  - Some are more about capacity (library, staging)
    - More sequential access
- Cost: Always an issue
  - On a fixed budget, I/O robs compute
  - Capability costs compute resources (more I/O nodes)

# Solution: Split it in two.

- Fast Scratch
  - 18x DDN SFA10000
  - 2,160 active 600GB SAS 15000 RPM disks
  - 36 OSS
  - InfiniBand QDR

- Long Term Fast Scratch
  - 8x DDN SFA10000
  - 2,240 active 2TB SATA 7200 RPM disks
  - 16 OSS
  - InfiniBand QDR



## Gaea filesystem architecture

FS and LTFS

# Implications

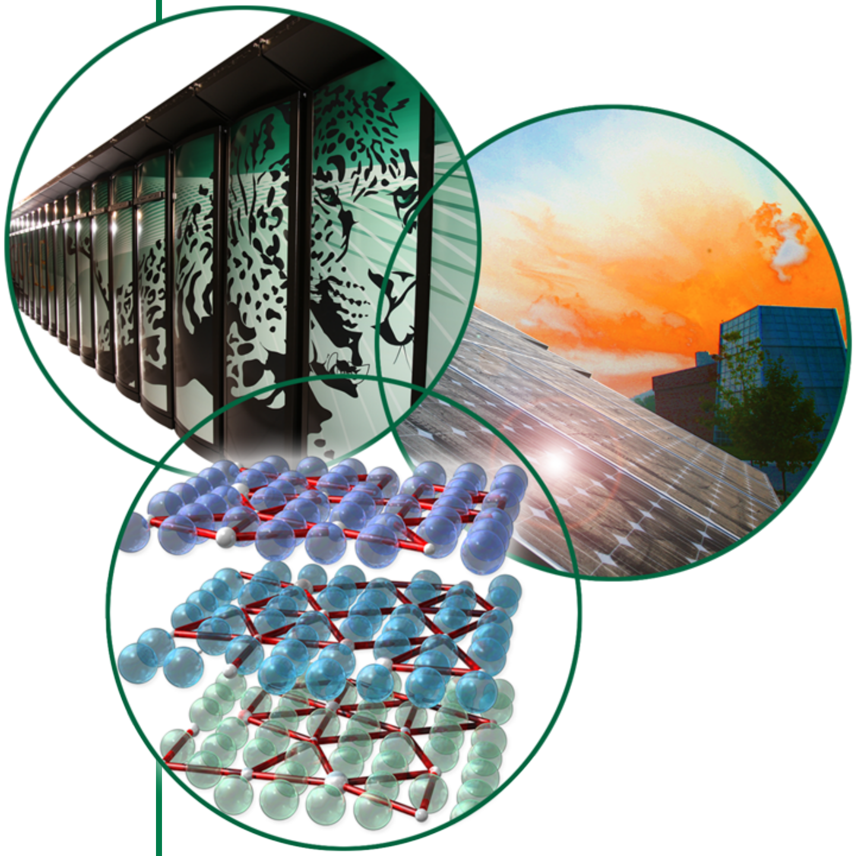
- Compute platform sees fast disk (FS)
  - Data staging is (hopefully) sequential and in the background
- Data staging done to bulk storage
  - Reduces cost for increased capacity
- Requires data staging step
- Leverage synergies where possible
  - Cross-connect between switches for redundancy
  - Combine data staging and some post-processing

# Data Movers

- Local data transfer nodes (LDTN)
  - 16x servers with dual InfiniBand
  - ORNL-developed staging parallel/distributed cp
  - Also handles simple post-processing duties
- Remote data transfer nodes (RDTN)
  - 8x servers with InfiniBand and 10Gb Ethernet
  - Wide area transfer with GridFTP
- Implies significant data handling overhead for workflow

# Thank You

Questions



# Lustre on Cray XT/XE Systems

- Individual compute nodes act as Lustre clients
- lnd for internal network (ptllnd, gnilnd)
- I/O nodes can serve as OSS nodes (“internal”) or route lnet to external network (“external”)
- External configuration used at NCRC
  - Improves flexibility
  - Enables availability to other systems