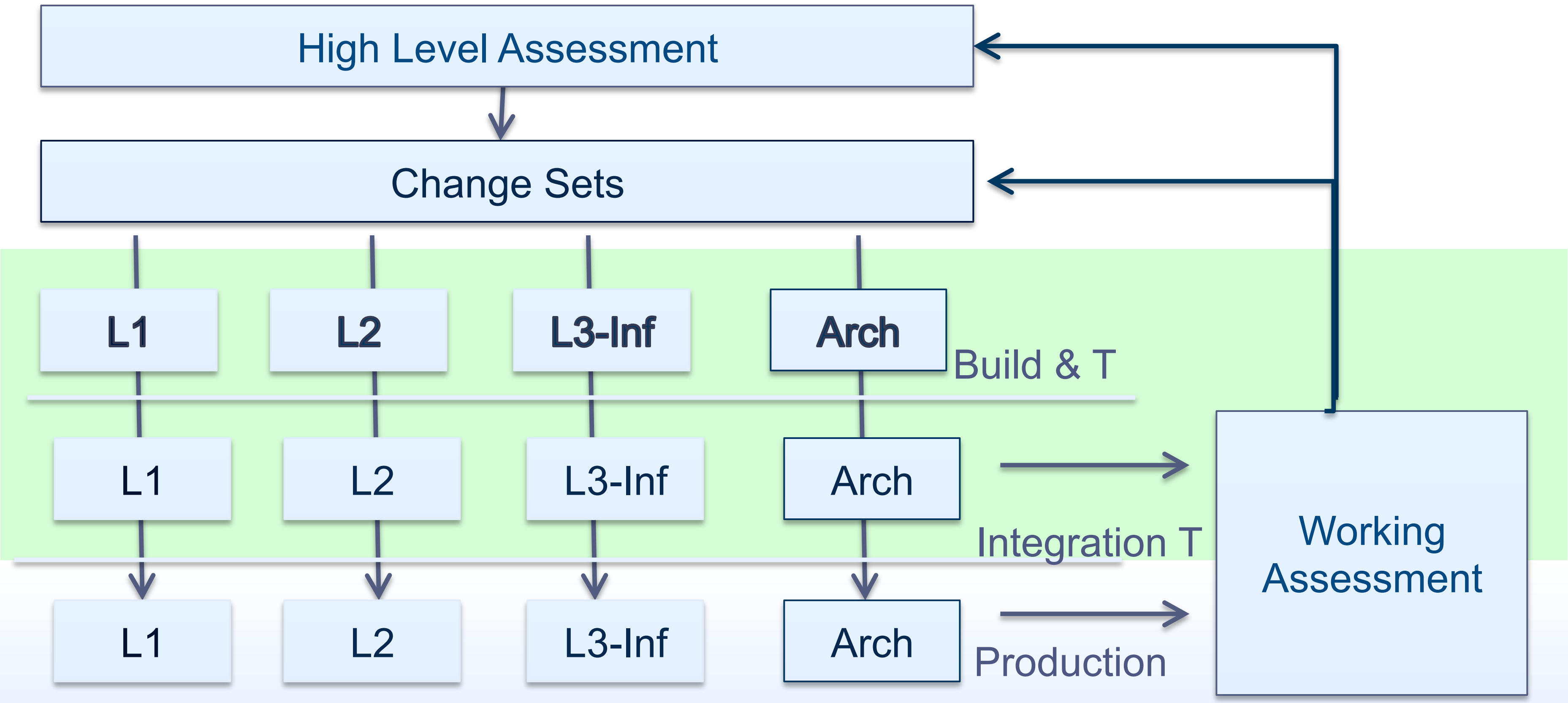
NCSA

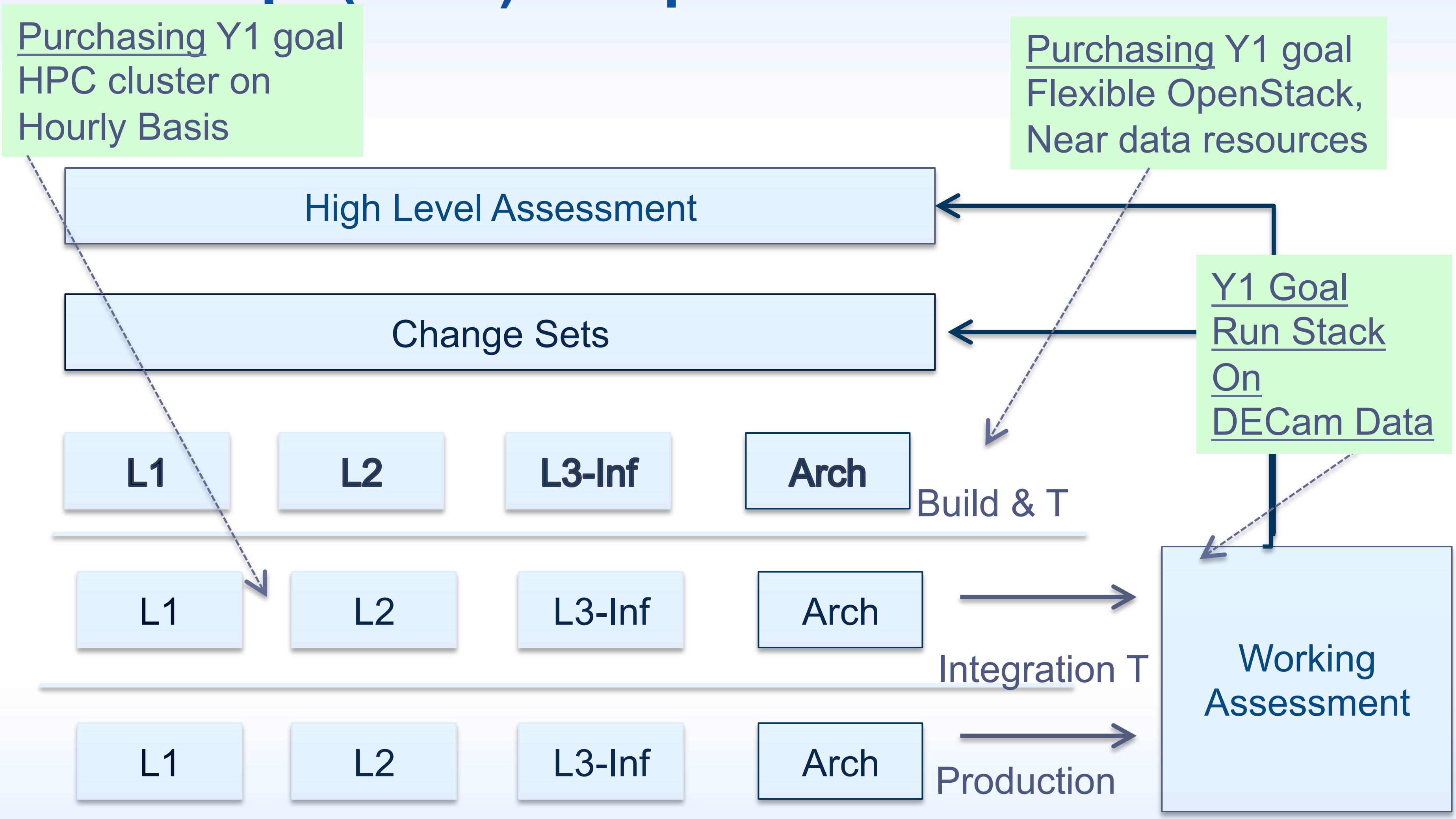
Lower-level DM architecture, including DM data challenges

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Concept (draft) of Operations



Concept (draft) of Operations

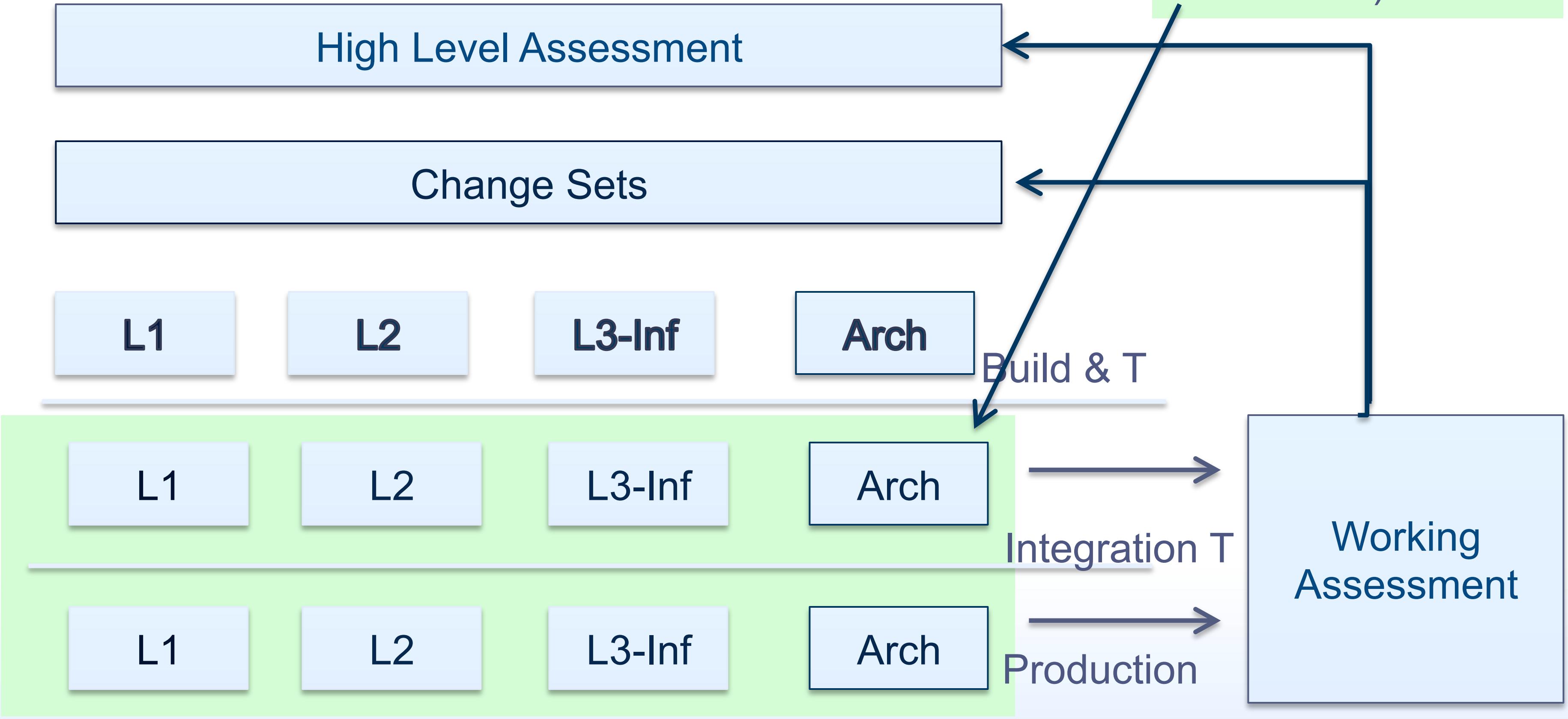


Devops – direction for infrastructure

- The LSST configuration system has embraced the DEVOPS model.
  - Clean separation of systems provisioning from application provisioning.
  - Consistent with good DES experience at NERSC.
  - Use of containers for production + software to manage the whole chain
    - We have people at a Velocity conference this week.
  - For testing, an NCSA is providing an OpenStack which will interoperate with development and test.
- NCSA assumes
  - Containerized deployments for production is a goal
  - However NCSA considers the final production infrastructure to be TBD, not necessarily related to OpenStack.
  - The project has a goal of making production infrastructure available generally.

Systems development work

Review strengths and weaknesses of implementation alternatives. (focus on storage and data movement)



High Level Assessment

Change Sets

L1 **L2** **L3-Inf** **Arch**

Build & T

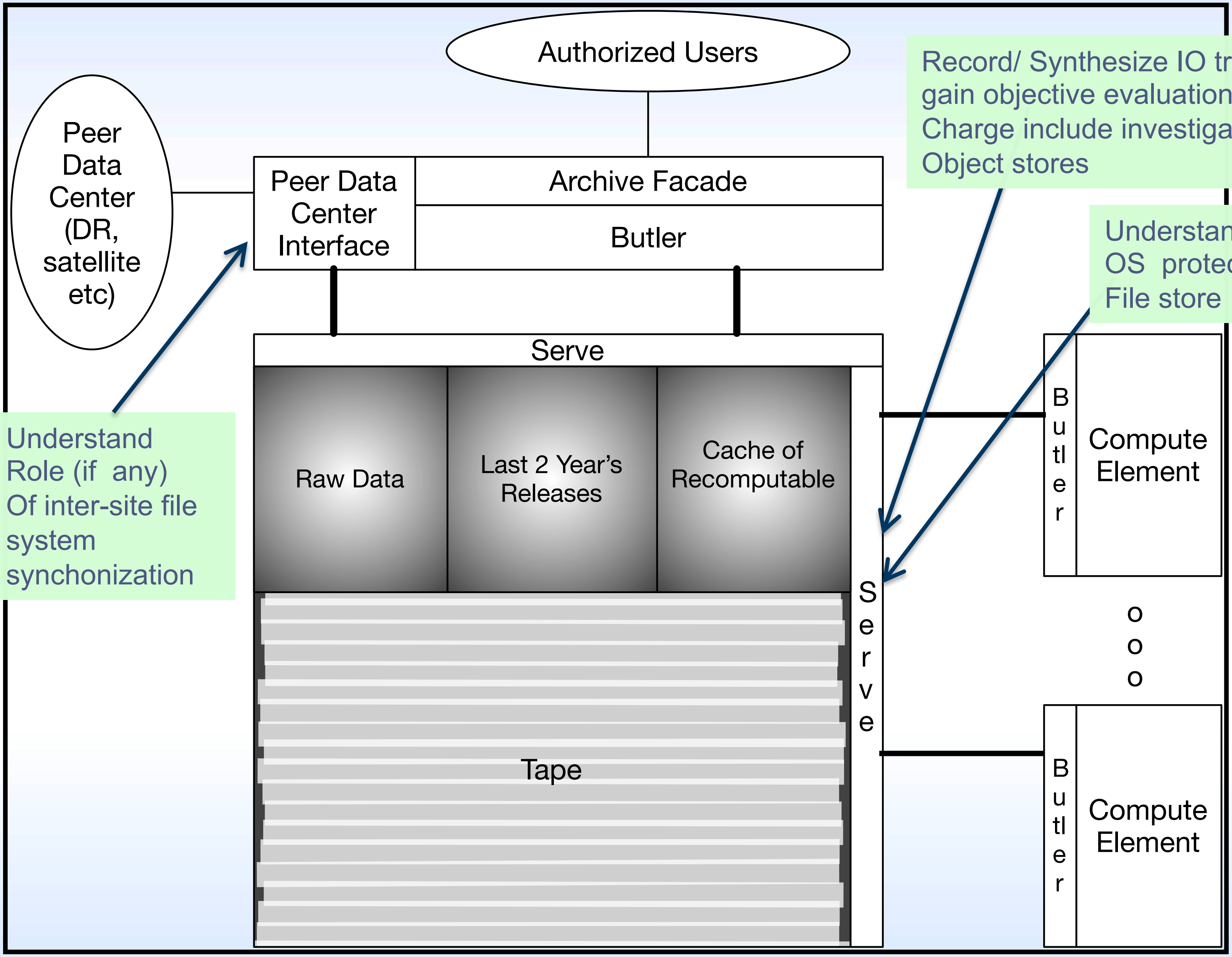
L1 L2 L3-Inf Arch

Integration T

L1 L2 L3-Inf Arch

Production

Working Assessment

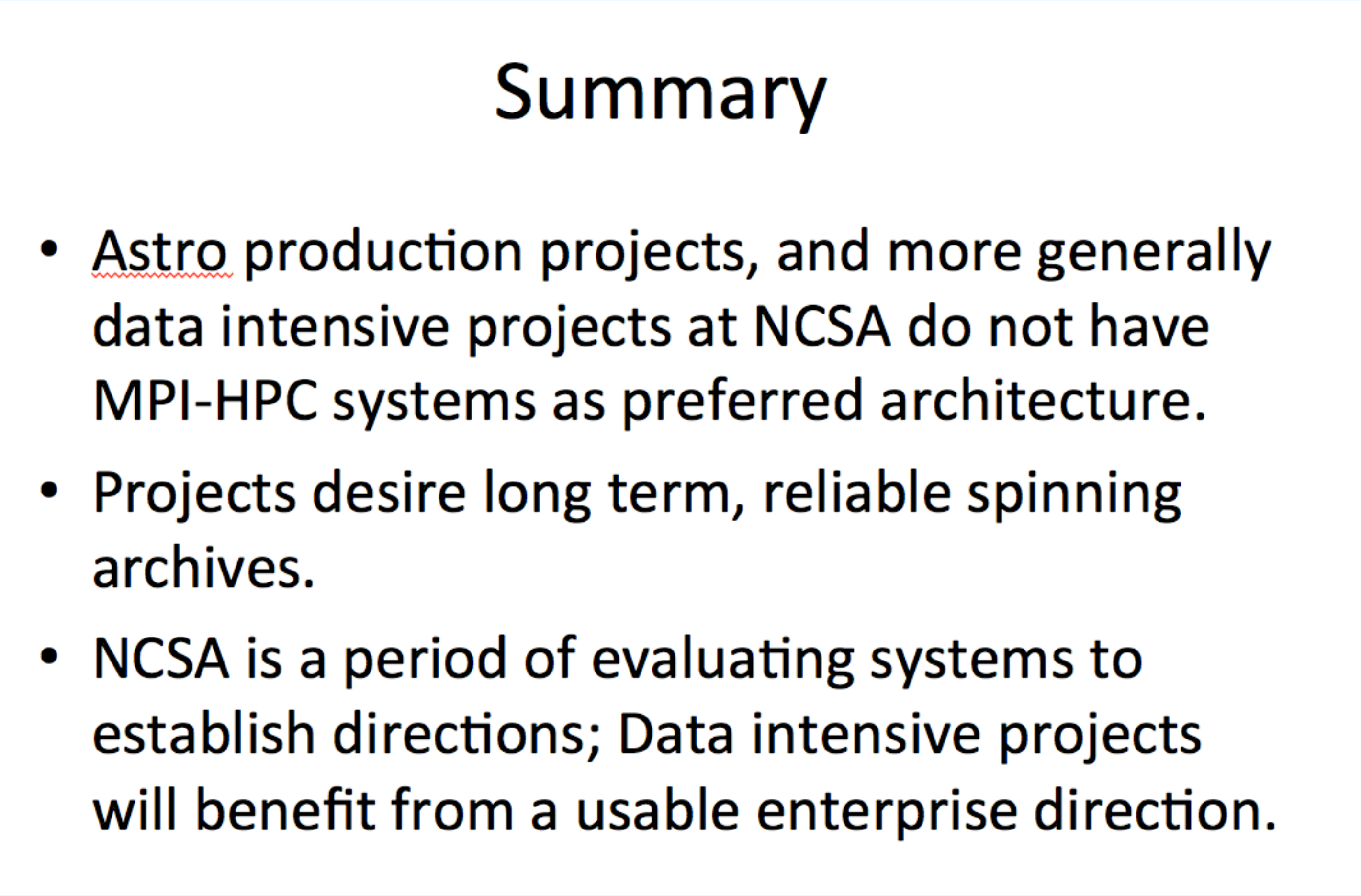


Record/ Synthesize IO traces to gain objective evaluation criteria Charge include investigating Object stores

Understand role of OS protections in File store

Understand Role (if any) Of inter-site file system synchronization

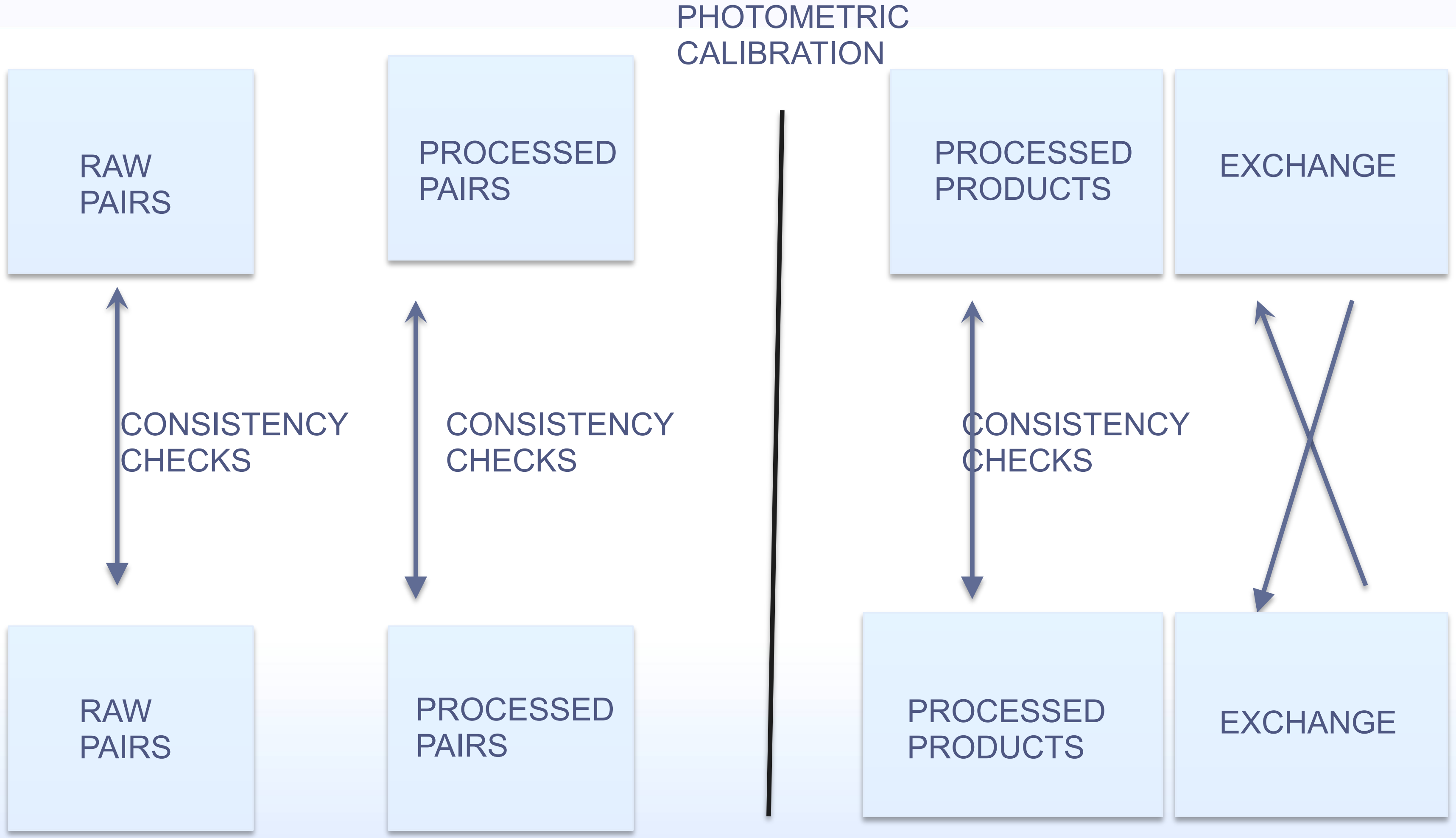
From another talk….



Summary

- Astro production projects, and more generally data intensive projects at NCSA do not have MPI-HPC systems as preferred architecture.
- Projects desire long term, reliable spinning archives.
- NCSA is a period of evaluating systems to establish directions; Data intensive projects will benefit from a usable enterprise direction.

Elements of a Level 2 Production Cadence



PHOTOMETRIC CALIBRATION

RAW PAIRS

PROCESSED PAIRS

PROCESSED PRODUCTS

EXCHANGE

CONSISTENCY CHECKS

CONSISTENCY CHECKS

CONSISTENCY CHECKS

RAW PAIRS

PROCESSED PAIRS

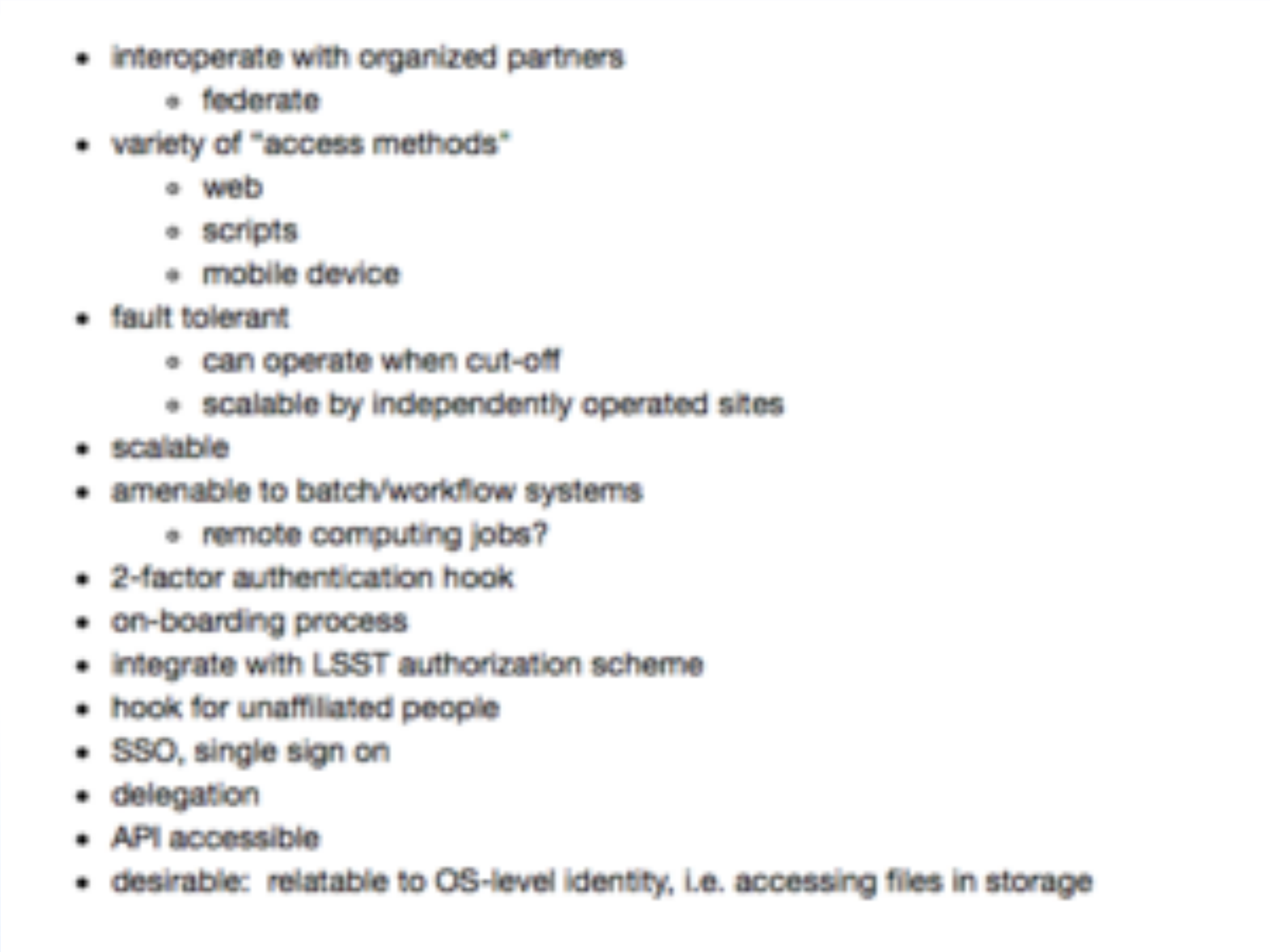
PROCESSED PRODUCTS

EXCHANGE

PROPOSAL in the air – LEVEL 2.5

- The project is proposing level 2.5 processing
  - Additional add-on processing based on competitive proposals from the community.
  - Would see the data concurrently with L2 processing.
  - Adds requirements to L2 processing.
    - Isolation of L2.5 products from non-essential L2 products.
    - Mutual Isolation of L2.5 products from each proposal.
  - TBD resolution of a number of process issues –
    - Early disclosure of L2 result to L2.5 PI's.
    - How I a L2.5 project to debug it's code and monitor its own state.
  - Initial proposal is L 2.5 runs at NCSA
    - No notional concept of operations yet.

L2.5 would interact with authentication/ Authorization system



- interoperate with organized partners
    - federate
- variety of "access methods"
    - web
    - scripts
    - mobile device
- fault tolerant
    - can operate when cut-off
    - scalable by independently operated sites
- scalable
- amenable to batch/workflow systems
    - remote computing jobs?
- 2-factor authentication hook
- on-boarding process
- integrate with LSST authorization scheme
- hook for unaffiliated people
- SSO, single sign on
- delegation
- API accessible
- desirable: relatable to OS-level identity, i.e. accessing files in storage

Workflow, Orchestration, Similar

- Past productions using Condor.
- Experiments with Pegasus up to scaling limit.
- Mechanisms?
  - Staging? Direct Access to file store?
  - Are all TBD.
- A consideration at NCSA is resource sharing with L2 production.
  - Large shared file systems may not have apropos uptime for L1.
  - Need to decide how coupled L1 and L2 are.

L2 Data Challenge Current Status

- The project has decided to cease data challenges for FY2015.
  - Enables a concentration on core software developments.
- Prior data challenges used national compute resources on Blue Waters and XSEDE.

Split DRP Summer 2013 Overview

Process Stripe82 ranges with an Overlap Region

- US/NCSA -40 < RA < 10 ~400 cores TACC Lonestar
- IN2P3 5 < RA < 55 ~700 cluster cores
- ~1,400,000 SDSS fields, ~2700 coadd patches for each team

DRP Stages/Tasks with Large Scale Parallelism

- Generate Calibrated Exposures – processSdssCcd
- Coaddition – makeCoaddTempExp,
  assembleCoadd, processCoadd
- Forced Photometry - forcedPhot
- Source Association - sourceAssoc
- Database Ingest of Results ~ 20,000,000,000 rows

Middleware Scalability Study Overview

Scaling Tests on TACC Lonestar

- Lustre parallel file system
- 1888 Dell M610 nodes, 12 cores/node
- Reference Input Data: SDSS fields
  - Identical Jobs process same 11 standard fields
  - Sample dataid “run=1033 filter=z camcol=2 field=12”

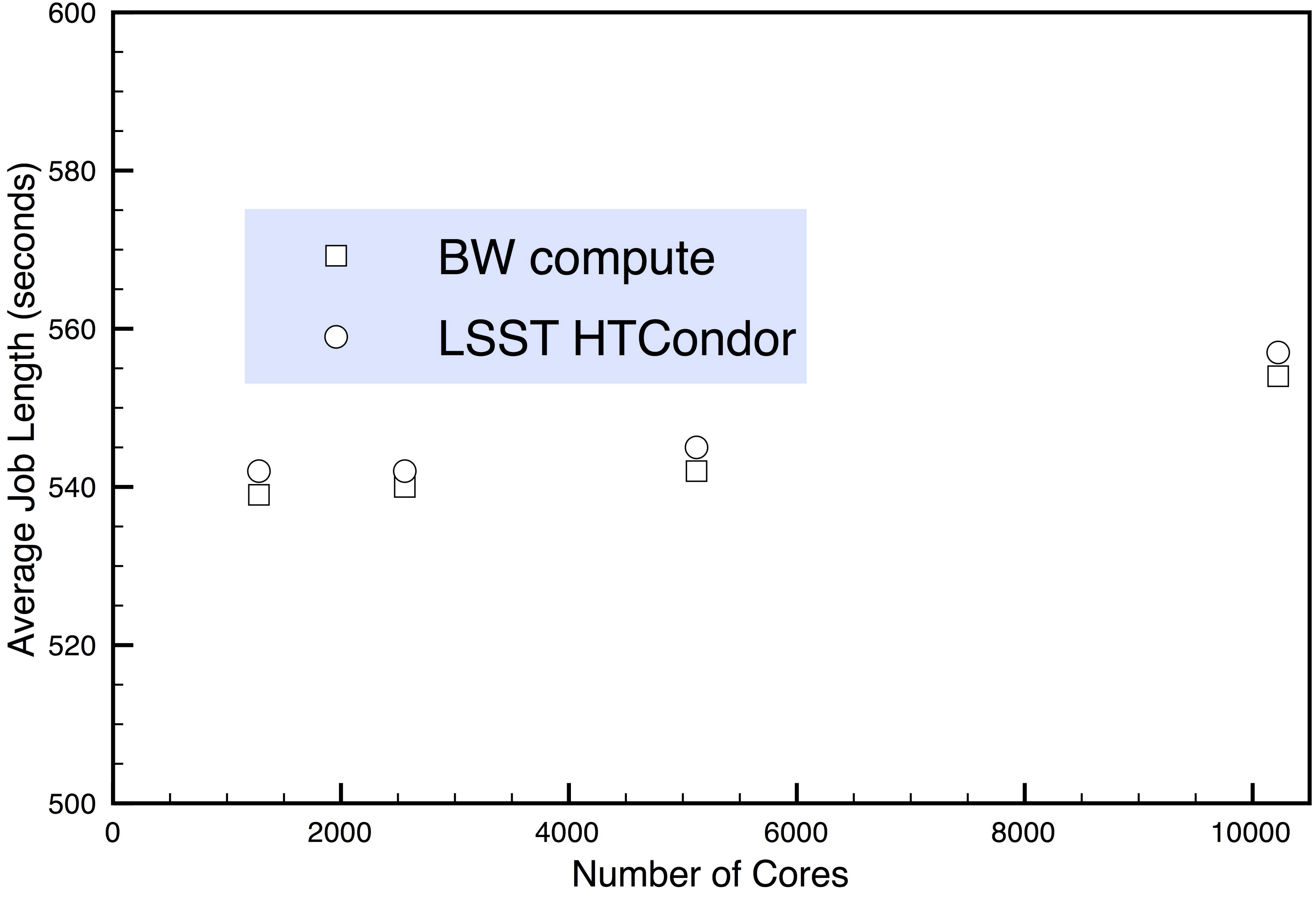
Run	Dims
B1	504 cores (42 nodes, 12 cores/node)
B2	1008 cores (84 nodes, 12 cores/node)
B3	2016 cores (168 nodes, 12 cores/node)
B4	4032 cores (336 nodes, 12 cores/node)

Middleware Scalability Study Overview

Scaling Tests on Blue Waters

- XE nodes: 2 AMD Interlagos 6276 CPUs, 16 cores/node
- LSST Software stack staged to local cache on compute node
- HTCondor GlideIn to LSST Central Manager
  - Application Launcher execs HTCondor on XE compute nodes
  - Condor Connection Broker for firewalled nodes
  - Multi-tier Collector on LSST Central Manager

Run	Dims
BW1	80 nodes/1280 cores
BW2	160 nodes/2560 cores
BW3	320 nodes/5120 cores
BW4	639 nodes/10224 cores


600
580
560
540
520
500
Average Job Length (seconds)
0
2000
4000
6000
8000
10000
Number of Cores
□ BW compute
○ LSST HTCondor