Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

Resources Preview

and review of terminology

Stu Fuess

Scientific Computing Portfolio Management Team (SC-PMT) Review

4 March 2015

Outline

- Hardware Resources
  - Processors
  - Disk
  - Tape
  - and a few other resources…

- In this introductory talk will give:
  - Terminology
  - Some background info
    - to “normalize” what you’ll hear in the presentations
  - Some hot issues
    - what to “wake up for”

Processors: Common Terms

- Glossary
  - The scientific computing discussed here is mostly performed on High Throughput Computing (HTC) farms (as opposed to HPC)
    - Farms is a deprecated term…
  - The farms are arranged as a computing Grid
    - Computing Jobs access Grid resources via Batch system

- Fermilab Grid resources (collectively FermiGrid)
  - CDF Grid (aka CAF) <<< going away soon
  - D0 Grid (aka CAB) <<< hope to go away
  - General Purpose Grid (aka GP GRID) <<< consolidating!
  - Private Cloud Workers (aka CW) <<< R&D

I'll show sizes later, but first: what metric is used?

Processors: Units of Measurement

- Our physical "unit" is one core with 2 GB physical memory and associated local disk space
  - Typical machines now are 64 cores, 128 GB ➔ 64 units
  - Year-to-year performance variation per core is not included
    - To ~30%, all cores equal
    - Jobs don't specify year/age of machine, so get a mix anyway
    - But we are moving toward normalized metrics (benchmarks)
      - e.g. CPU-hour on a 2.3 GHz AMD Opteron 6276
- Most past jobs fit within one "unit"
  - Same unit for Simulation / Reconstruction / Analysis jobs
  - The batch systems were organized similarly
    - One "unit" = One "core" = One "batch slot"

Processors: Jobs and Slots

- Job requirements are changing
  - We now see jobs requiring > 2 GB memory or ones that utilize multiple cores
    - Batch system evolving to allow jobs to request partitionable batch slots
      - Can ask for multiple processors, more memory
    - We are also preparing for different HW architectures to appear
      - Job may need access to many-core systems: GPU, etc…
        - Small test cluster with GPUs, Intel Phi – but nothing for production
- There's an evolution of how we manage resource requests and accounting
  - Move from "slot" concept to "CPU-hours"
  - Some multi-unit resource requests will need multiplicative factor

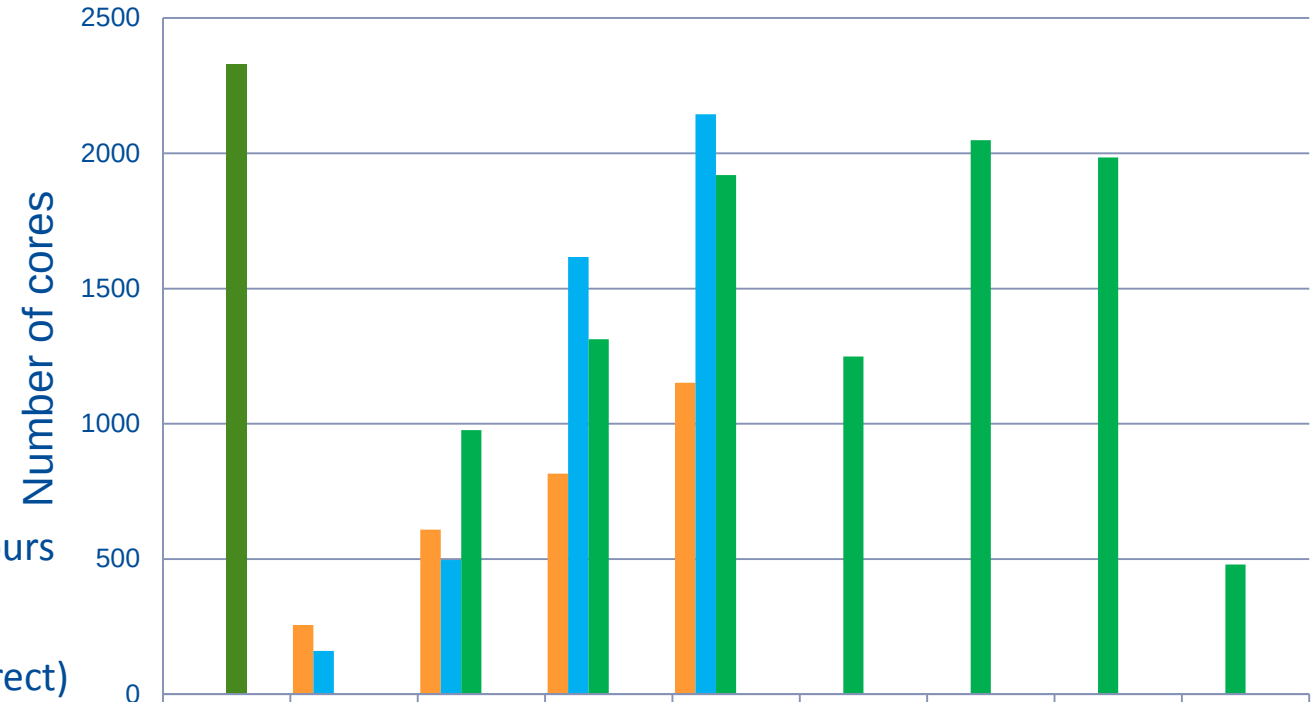
Processors: Current GRID Configurations

Grid	Cores
CDF	2832
D0	4416
GP	9968
CW	2328

Equivalent to ~145M CPU-hours in one year

Replacement cost ~ $2M (direct)

3/2015 status: Cores vs Year Purchased



	FY07	FY08	FY09	FY10	FY11	FY12	FY13	FY14	FY15a
CDF Cores		256	608	816	1152				
D0 Cores		160	496	1616	2144				
GP Cores			976	1312	1920	1248	2048	1984	480
CW Cores	2328								
Years of Warranty	3	3	3	3	4	5	5	5	5
Wty. thru	2010	2011	2012	2013	2015	2017	2018	2019	2019

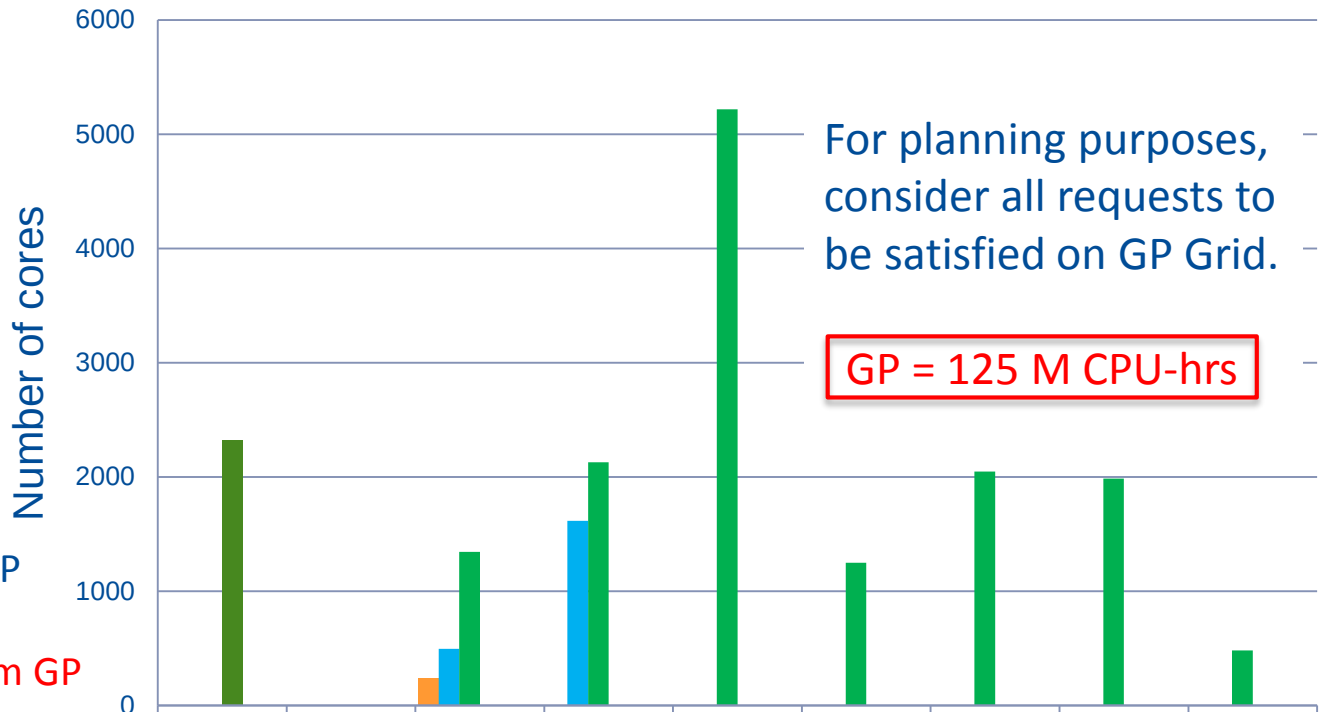
Processors: Plan for remainder FY15

Grid	Cores
CDF	240
D0	2112
GP	14448
CW	2328

Move CDF, D0 resources to GP

Requests will be satisfied from GP

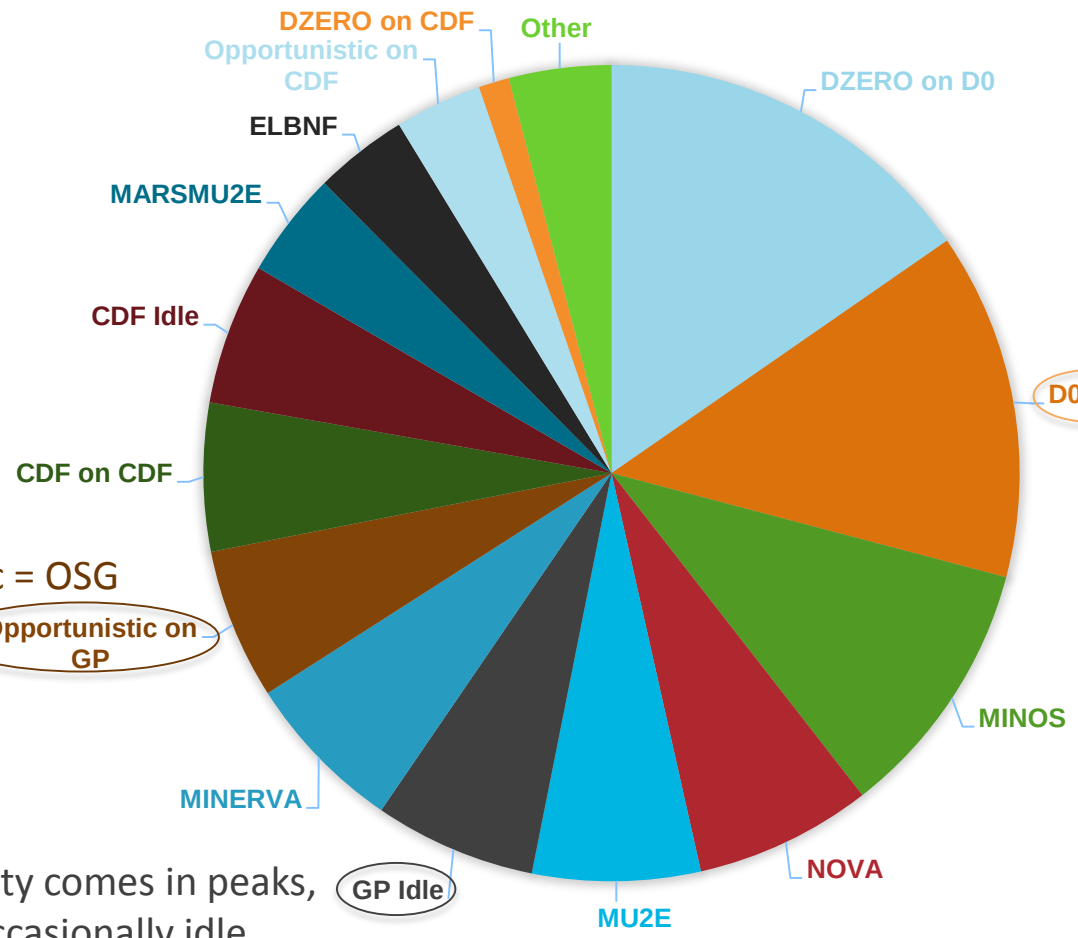
FY15 Plan : Core vs Year Purchased



	FY07	FY08	FY09	FY10	FY11	FY12	FY13	FY14	FY15a
CDF Cores			240						
D0 Cores			496	1616					
GP Cores			1344	2128	5216	1248	2048	1984	480
CW Cores	2328								
Years of Warranty	3	3	3	3	4	5	5	5	5
Wty. thru	2010	2011	2012	2013	2015	2017	2018	2019	2019

Processors: Past Usage (Mar 2014 – Feb 2015)

Total ~145M CPU-hours

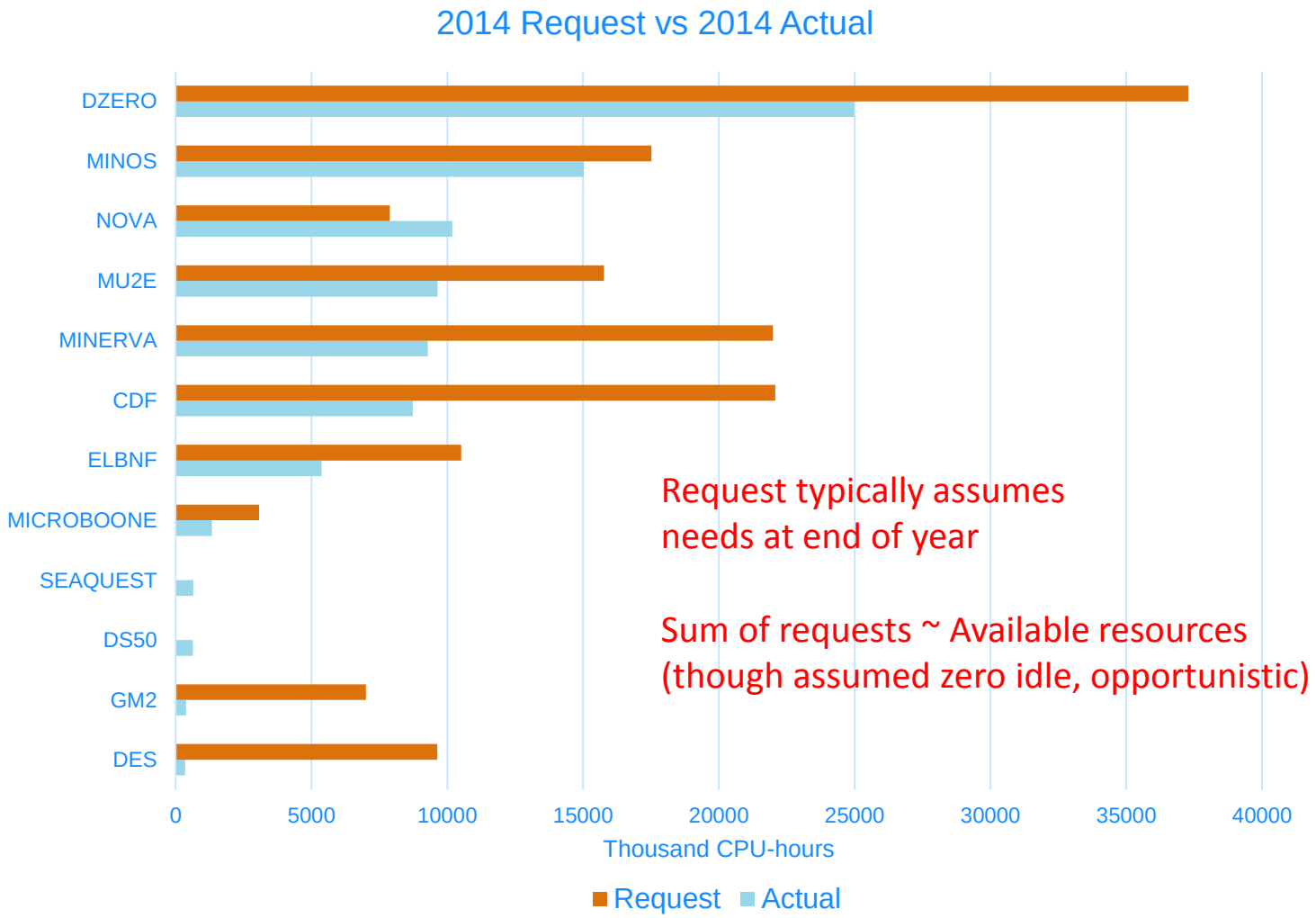


A lot of D0 resources unusable by others

Opportunistic = OSG

If activity comes in peaks, then occasionally idle

Processors: Last Year's Request vs Usage



Processors: Other Resources

- Computing activity often comes in bursts
  - New pass over old data
  - Conference / paper preparations
- If Fermilab facility sized for average utilization, then there will be times when peak demand cannot be (immediately) satisfied
- Alternatives:
  - Most experiments can submit jobs to the OSG
    - Opportunistic use at other sites
      - Lower probability of job running to completion before eviction
      - Data will need to be accessed across the network, so slower
      - Free!
    - Paid Cloud services (e.g. Amazon, MS Azure)
      - We are working on a "virtual facility" – but complex economic model

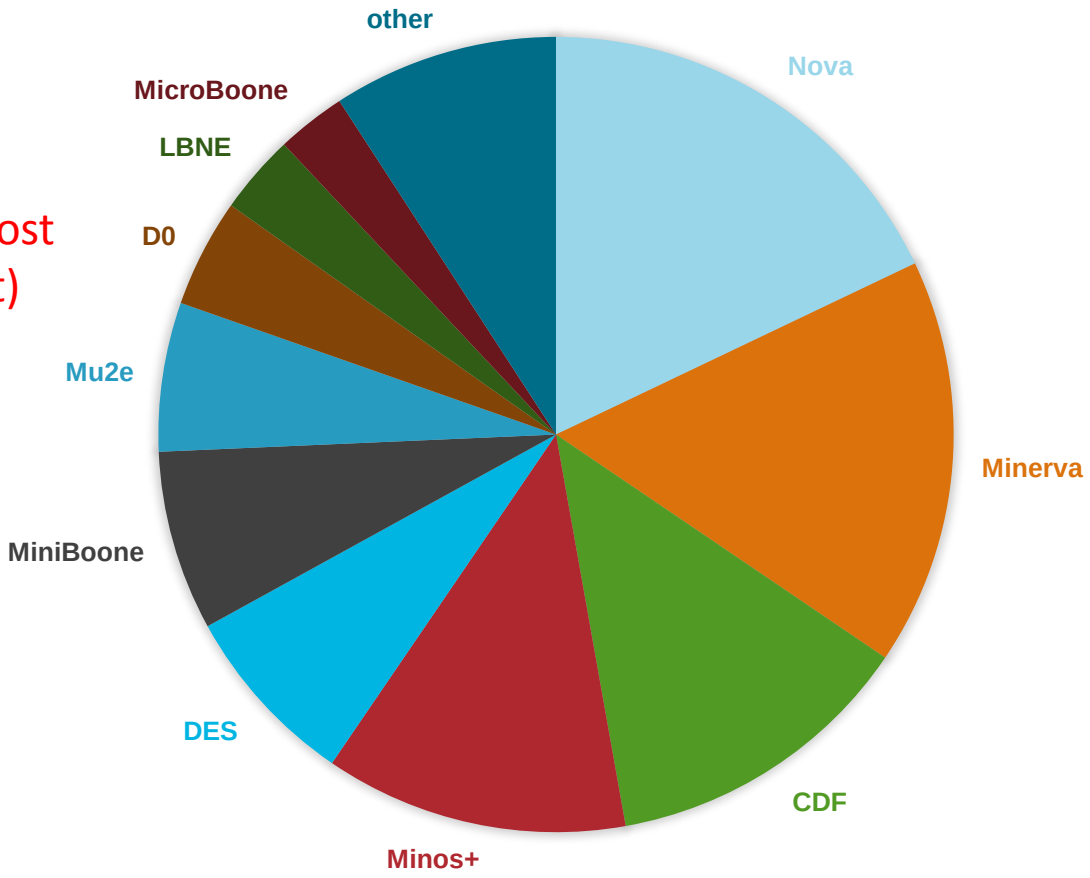
Disks: Common Terms

- Scientific Computing uses disk space on:
  - BlueArc
    - High performance, high cost ($1000/TB) NAS operated by CCD
    - POSIX compatible (can do all I/O operations) on local systems that mount via NFS. What people are used to seeing!
    - Not directly accessible off-site
    - Can be overloaded => no longer mount on Grid nodes
    - Easiest to use for development, local analysis
    - Allocations per experiment; nearly always full
  - dCache
    - Highly distributed storage with central name space
    - Much lower cost ($100/TB)
    - Read / Write interfaces, but does not look like usual file systems
    - Accessible from off-site
    - A cache (optionally front-end to tape system), so old files are flushed

Disks: Major BlueArc Users

Total ~1.9 PB

Replacement cost
~ $2M (direct)

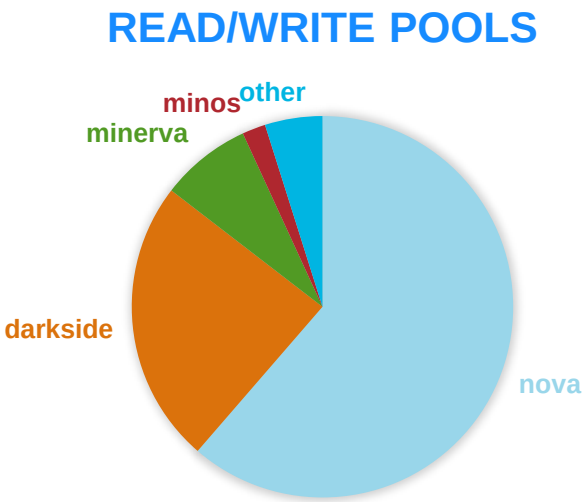


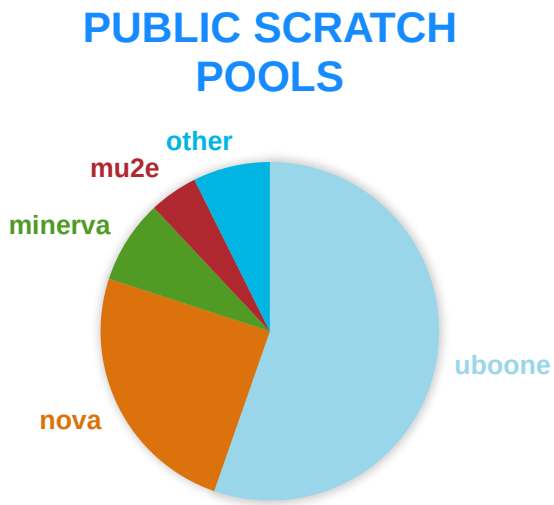
Disks: dCache Resources

Pool Type	Sum of Total TB	
readWritePools	2442.4	Tape backed
PublicScratchPools	656.6	Not tape backed
SfaPoolGroup	75.8	Internal use by tape facility
MinervaWritePools	65.5	Raw data from experiment DAQ
ArchivePools	59.9	Used by Active Archive Facility
DESPools	45.1	DES
NovaWritePools	39.9	Raw data from experiment DAQ
LQCDPools	39.9	LQCD to/from tape
RawDataWritePools	24.1	Raw data from experiment DAQ
EmptyPools	22.1	Currently not in use
ExpDbWritePools	22.1	Database backups
FermigridVolPools	9.4	FermiGrid scratch area
Grand Total	**3502.9**	

3.5 Petabytes total
~ $100/TB (direct cost)
~ 3x for surroundings
→ ~ $1M investment

Disks: Major dCache Users





2385 TB

640TB

How do you interpret requests for cache disk space?

- Transient space for writing raw or reconstructed data to tape
- Transient space for staging data for reconstruction / analysis passes
- Long-lived space for data that will be accessed multiple times

Note that “permanent” space not available in current model for dCache

- But working on design to provide this functionality

Tape

- Current complex has 7x 10,000 slot tape libraries
  - 3x are dedicated to CMS
  - Remaining 4x, with most recent tape technology (> 8 TB per) provides ~320 PB of capacity
    - So capacity is not an issue

- Assume that ~ $50/TB for tape media and peripherals
  - So cost is only an issue when speaking PB size volumes

- Will track requests and provision appropriately, but otherwise need not pay much attention to tape storage requests

Other resources

- Presentations may also mention:
  - GPCF static VMs...
    - Virtual Machines, used for interactive login, special services
  - cvmfs
    - A software tool for widely distributing a file system (typically containing the experiments code and applications) and make it look like a locally mounted file system
    - Requires associated repositories and distribution servers
  - Build and Release service
    - A software tool (Jenkins) and associated hardware for making frequent software builds and package releases on different platforms

Backup slides

The partial year problem

- Processing needs are analyzed in terms of the total number of CPU-hours requested over a year duration
  - and hence converted into the number of machines needed
- Some requests (notably Mu2e) are made relative to CPU-hours needed for the remainder of FY15 (7 months, Mar-Sep)
  - For the purpose of the analysis, these requests are converted to the annual equivalent, i.e. multiply by 12/7