We will be having our regular bi-weekly Infrastructure WG telecon on Monday, May 10, at 12N CT (10A PT).

Agenda

- **Yale Patt** talk on Apr 30 at Illinois
  - "Future Microprocessors: Multi-core, Mega-nonsense, and What We Must Do Differently Moving Forward"
  - Good talk (highly recommend), but main point for LSST: Microprocessors of the future will be a hybrid combination of traditional CPU cores and GPU/SIMD "cores".

- Shared Memory Architectures
  - **ScaleMP**
    - Notes here -> InfrastructureWGMeetingD20100426
    - debrief
    - [Arun] vSMP is used at SDSC on its pre-Gordon machine called DASH. I am actively involved on this effort also.
    - [Arun] Regarding experiences, I interact with them on the high-level technical issues. We could discuss this over the phone on one of our Infra-WG calls.
  - [Arun] In addition to scaleMP you might also want to get information from ?http://www.3leafsystems.com/. They are taking a h/w cum s/w solution rather than the fully s/w solution approach by scaleMP. I can put you in touch with them..."
  - **Next steps?**
  - SGI Altix; 1536 cores; 8TB memory
    - ?http://www.ncsa.illinois.edu/News/10/0302NCSAprovide.html
  - New Shared Memory HPC Machine at NCSA (Ember)

- TeraGrid TRAC Allocation
  - TeraGrid? Award - 1.51M SUs, 400TB tape, 20TB disk (Apr 1, 2010 to Mar 30, 2011)
    - The **20TB is on GPFS-WAN**; gridftp-able
  - Have **5TB on NCSA Lustre**; maybe get 5TB more, but probably not 15TB more
  - Moving 1000 abe SUs to **1108 Cobalt SUs** for MOPS
  - Creating TG gateway and community accounts (lsst, lsstread) for accessing mass storage
  - Adding/updating **Authorized LSST members** from the startup allocation to the new HPC allocation
  - Meeting today with HPC filesystem staff to **get a better estimate for filesystem bandwidth** -- Information has come to light that our previous estimate of ~140MB/s between abe and the spinning disk (scratch and project space on able) is low.

- Existing Resource Usage Update (as of Mar29)
  - TeraGrid Resources (Startup Allocation)
    - Service Units
      - Abe: Allocated: 30K SUs; Remaining ~29.4K SUs
      - Lincoln (nVidia Tesla GPUs): Allocated: 30K SUs; Remaining 30K SUs
        - **can "convert" some SUs to lincoln for GPU testing later**
    - Disk Storage
      - Allocated: 5TB; Remaining: 5TB
    - Tape Storage
      - Allocated: 40TB; Remaining: 40TB

◊ Authorized Users: GregD, DavidG, K-T, SteveP, RayP, Jonathan Myers

- DC3b Infrastructure for the Performance Tests
  - ♦ ?http://dev.lsstcorp.org/trac/wiki/DC3bHardwareRequirements
  - ♦ ?LSST-11 DC3b Hardware
  - ♦ ?https://www.lsstcorp.org/docushare/dsweb/Get/Document-8529/LSST-TeraGrid-Proposal.pdf
  - ♦ Compute
    - ◊ We're good on this. 1.51M SUs awarded from TG.
  - ♦ HPC Project Space
    - ◊ We're good on this. 20TB awarded from TG.
  - ♦ Database Disk
    - ◊ We're good on this. 15TB for database on lsst10 as of Feb 22, which covers the requirements for all of DC3b.
  - ♦ Tape Storage
    - ◊ We're good on this. 400TB of dual copy mass storage awarded from TG. No tape gap. No need to purchase additional tapes. No data loss is expected due to media failures.
  - ♦ **Database Ingest** (Jacek)
    - ◊ ?http://dev.lsstcorp.org/trac/wiki/DC3bDbIngest
    - ◊ ?http://dev.lsstcorp.org/trac/wiki/dbPartitioner
    - ◊ New requirement? Temporary "scratch databases"; two parts:
      - · a database, likely centrally located but potentially per-node
      - · TSV files
      - · Is any of this DC3b?

- DC3b User Access
  - ♦ DC3bUserAccess
  - ♦ Unique Identifier for Logical Set of Related Files
    - ◊ discussion w RHL -- pending feedback from him
    - ◊ DC3bUserAccess (this is what we're talking about, but don't look at it yet)
  - ♦ Bulk Upload Into Catalog
    - ◊ DC3bUserAccess
    - ◊ assuming standard mysql utilities
    - ◊ assuming storage requirements are not significant
      - · MikeF proposing stmt re user expectations for storage
  - ♦ Web Data Server update
    - ◊ DC3bUserAccess
    - ◊ ?http://jira.ncsa.uiuc.edu/browse/LSST-106
    - ◊ **Will be installed/configured after lsst2 (iRODS server) is up and running**
  - ♦ Image Cutout Service update (K-T)
    - ◊ DC3bUserAccess
  - ♦ Sample Scripts
    - ◊ IPAC (Suzy)
  - ♦ Web Interface
    - ◊ IPAC owns this
    - ◊ IPAC (Suzy); interface to scripts; reuse existing portals
    - ◊ **Briefing from the Gator discussion** on March 23 (Jacek)
  - ♦ **Database Server(s) at SLAC** (Jacek)
    - ◊ 2 servers; qserv; 15-20TB
    - ◊ expected to be ready before May
    - ◊ Apr12 status: still on track for having the secondary database server(s) ready by May1

- Database replication strategy (Jacek)
    - dbBackupRepl?

- DC3b User Support
    - IPAC owns this
    - Separate item from above
        - User Access is about systems and software; User Support is about receiving questions/problems from human beings
    - Active discussion going on among SuzyD, DickS, MikeF
    - One line summary: Ticket system would be good, KB would be good, no labor resources available, planning on an email address, discussions continue
    - Support email address: dc-support at lsst.org
        - Scope: user support for the data challenges
        - support at lsst.org is too generic
        - all dc issues -- do not try to have user select "category" of issue
    - Bring in Ephibian?
        - both recommendations & implementation

- **ImSim Data Management with iRODS** Update (Arun)
    - Sites: UWash, Purdue, SLAC, NCSA
    - ImSimDataManagement
    - DavidG is getting lsst2 ready as the iRODS server at NCSA
    - Apr12 status:
        - Some hardware and firewall issues at Purdue
        - **Expect data flowing by the end of this week (i.e. Apr16)**

- **Output Data Management with REDDnet** Update (MikeF)
    - ?http://docs.google.com/View?id=dgvmjj2x_16f4mvfmd6
    - 2x24TB (48TB) going to both NCSA and SLAC; depots exist at Caltech and elsewhere
        - **Update: Will be shipped by the end of next week (May 7)**
    - big focus on monitoring by team at Vandy
        - perfSONAR suite (I2) (snmp), BWCTL (iperf), MRTG (snmp), Nagios, and custom
        - monitors availability, throughput, latency, general health, alerts
    - single virtual directory structure -- sandbox for lsst created
    - L-Store
        - provides client view / interfaces (get/put/list/etc.)
        - defines the virtual directory structure
    - StorCore
        - partitions REDDnet space into logical volumes (think: LVM)
        - L-Store uses StorCore for resource discovery
    - Web interfaces for both StorCore and L-Store
    - Example code available (contact mike to get a copy)
        - upload.sh file1 file2 dir1 dir2 remotefolder
        - download.sh removefile1 remotefile2 localdestination
        - ls.sh remotefile or remotedirectory
        - mkdir.sh remotedir1
        - additional commands to "stage" files across depots
    - **on schedule to have new servers at sites before PT1 starts**

- Update on LSST Database Performance Tests Using SSDs (Arun/Jacek?)

- LSST expects to manage some 50 billion (50*109) objects and 150 trillion (150*1012) detections of these objects generated over the lifetime of the survey. This data will be managed through a database. The current baseline system consists of off-the-shelf open source database servers (MySQL) with a custom code on top, all running in a shared-nothing MPP architecture.
- To date, we have run numerous tests with MySQL to project performance of the query load we expect to see on the production LSST system, including low volume, high volume and super high volume queries (simple queries, full table scans and correlations, respectively). Based on these tests we estimated hardware needed to support expected load. All these tests were done using spinning disks.
- Having the opportunity to redo these tests with solid-state technology (solid state disks, or SSD) would allow us to understand potential savings and determine whether SSD could help us simplify the overall architecture of the system by approaching things in a ?different? way than on spinning disk.
- The tests we expect to run include:
    - Selecting small amount of data from a very large table via clustered and non-clustered index (this is related to low volume queries).
    - Verifying whether we can achieve speed improvements for full table scans comparable to raw disk speed improvements seen when switching from spinning disk to SSD (this is related to high volume queries).
    - Testing architecture that involves heavy use of indexes, including composite indexes instead of full table scans for high volume queries.
    - Executing near neighbor using indexes on subChunkId without explicit subpartitioning.
- We expect to run these tests using USNOB data set, which, including indexes and other overheads fits on ~200 GB.
- **Status**: waiting on accounts

- Update on Lawerence Livermore database scalability testing (DanielW)
    - Description: LLNL has provided a number of nodes (currently 25) as a testbed for our scalable query processing system. Being able to test over many nodes allows us to understand where our query parallelism model succeeds and fails, and helps us develop a prototype that can handle LSST database query needs. So far, use of this many-node cluster has uncovered problems in scalability in job control, threading, messaging overhead, and queuing, which we have been incrementally addressing in each new iteration (3 so far).
    - **Status**: developing and testing a new model since tests in Jan showed bottlenecks at >4 nodes
    - Hoping to get time on a 64 node cluster at SLAC
    - software will be installed on lsst10 after testing
    - [Jacek] New Resource: A 64-node cluster at slac (used to be for PetaCache tests), which we will be able to use for lsst related scalability tests (kind of permanently). Total of 128 CPUs, 1 TB of memory (16 GB per node), 2 TB of total local storage (34 GB per node).

- Server Administration at NCSA
    - With the upcoming DC3b runs and the increased need for system reliability with introduction of end user access to our DC data, we're tightening up the processes and procedures related to the administration of the LSST servers at NCSA
    - New email address: lsst-admin at ncsa.uiuc.edu
        - Scope: technical issues, questions, problems with the servers located at NCSA
    - Define Roles & Responsibilities
    - **prep work for DC3b ongoing** (mainly lsst10 and buildbot servers)
    - **qserv to be installed on lsst10 after PT1**

- Directory Structure for Image Repositories
    - [DC3bDataOrganization]
    - ?http://jira.ncsa.uiuc.edu/browse/LSST-96
    - **directory structure firming up**
    - **still need to settle on "tarring" strategy** for mass storage

- Getting input data ready for PT1 runs
    - moving CFHTLS data to mss
    - moving ImSim data to mss (see iRODS above)

- **Using GPUs to Accelerate Database Queries**
    - [TimA] "I just ran across Accelerating SQL Database Operations on a GPU with CUDA, which is the first application I've seen of GPUs to SQL. I haven't read it carefully yet, but a quick skim suggests that they transfer tables of a few million rows into the GPU memory, transforming them into column form on the way. The model is that repeated SELECTs are done on these tables, so that the transfer time is unimportant. Optimistic, no doubt, but even including the transfer time they get speedups over 20X."
    - ?http://www.cs.virginia.edu/%7Eskadron/Papers/bakkum_sqlite_gpgpu10.pdf

- Building the Data Center of the Future 2nd Biennial Workshop: HPC Data Centers
    - Jun 23-24, 2010. Champaign, IL.
    - ?http://www.ncsa.illinois.edu/Conferences/DataCenter/

- Cost Sheet Update
    - Baseline version is v45
        - ?https://www.lsstcorp.org/docushare/dsweb/Get/Version-12185/Infrastructure-Costs-v45.xls
        - caveats apply: v45 does not *exactly* match PMCS
    - Current version now v74
        - ?https://www.lsstcorp.org/docushare/dsweb/Get/Document-6284/Infrastructure-Costs-v74.xls
        - ?https://www.lsstcorp.org/docushare/dsweb/Get/Document-8189/CostSummaryWithBaseline-
    - Summary of Changes
        - xxx
    - Upcoming Changes
        - Priority is updating the Power & Cooling estimates
            - ?LSST-10 Update Power & Cooling at Base Site (info already received from RonL)
            - ?LSST-47 Power Costs at BaseSite: Use Historical Data to Model Future Power Prices
            - ?LSST-36 Update Power & Cooling at ArchSite
            - ?LSST-36 P&C and Floorspace at PCF (rates, payment approach, green features of PCF)
        - ?LSST-78 Move the 3% CPU spare from document 2116 "CPU Sizing" to document 6284 "Cost Estimate"
        - ?LSST-79 Add tape library replacement to ArchAOS and BaseAOS
        - ?LSST-28 Optimal CPU Replacement Policy
        - ?LSST-14 Processor Sizing Update (Doc2116 LSST CPU Sizing)
        - ?LSST-37 Missing controller costs for disk
    - Next steps with cost sheet
        - Full review each of the elements of the cost sheet (boxes of the mapping document)
            - More readable description of the formulas being used

            - Identification and documentation of assumptions
            - Identification and documentation of external data input
        - Serves two significant purposes
            - Allows for better internal reviews (validation of models and information used)
            - Provides justifications for external reviews
        - Results in an updated (or replacement of) Document-1684 and related documents ("Explanation of Cost Estimates")

- InfraWG Ticket Update

Notes

Attendees: KTL, JacekB, BillB, JeffK, ArunJ, MikeF

- updates on DC3b-related
    - TG allocation status
        - ready is all respects, no known action items remaining
    - iRODS/ImSim status
        - Arun gave update
        - 2TB of space available on lsst2; no stoppers; mss not in critical path
        - will be done by the end of the week (May 14)
    - calipso data access
        - completed
- recent meeting with ScaleMP
    - Mike in contact with vendor
    - discussed Dash (which use vSMP)
- REDDnet update
    - servers in transit to NCSA, SLAC
- recent Yale Patt talk
    - future chips will be a combination of GPU cores and SIMD cores
- (potential topic) database backups
    - Jacek working on recommendations; not yet cooked enough for infra team
- Jacek and Mike established basic guidelines for db admin roles
    - [DatabaseAdminRoles]
- Debrief on lsst9 memory problems last week

Useful Links

- InfraWG Home Page
    - ?http://dev.lsstcorp.org/trac/wiki/InfrastructureWG
- InfraWG Tickets (in priority order)
    - ?All InfraWG Tickets