
Node a01
00:E0:81:2A:B3:0D
Node a02
Node a03
00:E0:81:2A:F2:65
Node a04
00:E0:81:2A:F3:A5
Node a05
00:E0:81:2A:F3:17

SGE & LSF

Hello!

- I'm Chris
  - 'dag@sonsorol.org' (public)
  - 'chris@bioteam.net' (corporate)
- I'm the only thing keeping you from dinner & drinks upstairs
- I work for the BioTeam
  - http://bioteam.net
  - Independent consultant shop
  - Scientists self-taught at IT
  - Bridging the science-HPC gap
    - Life science
    - Oil & Gas
    - Government
    - Digital Content Creation

Bias Disclosure

- I'm an industry suit
  - Cynical
  - Focused on practical, deployable solutions
- In case you think I'm a shill:
  - http://bioperl.org
  - http://gridengine.info
  - http://xml-qstat.org

Topics

- Why this talk?
- Enterprise DRM “Case Study”
- New in LSF 7.x
- More details from “Case Study”

Why this talk?

Original Title:

"*Confessions of a SGE zealot: Why I used Platform LSF on my last large project*"

This talk is NOT …

- An attack on SGE
- A deep technical comparison
- A marketing/sales “competitive positioning” presentation

This talk IS …

- About how one large enterprise selected, purchased and uses DRM technology
- Specifically about how client requirements led to the selection of Platform LSF
- A bit about the larger “SGE vs. LSF” differences and similarities

- Note: Your experiences *will* vary. These are personal thoughts and observations

If you have downloaded this

- You do not have my permission to redistribute or excerpt *parts* of this presentation
  - All or nothing
- You do not have my permission to quote me selectively or use these materials for marketing
- *My career depends on objectivity & independence*

Enterprise DRM “Case Study”

Enterprise DRM “Case Study”

- Fortune 20 multinational
- New construction
- ~ $300M Facility
- Hundreds of researchers
- East Coast, USA
  - Varied Science
    - Chem/Bio/Genetics
    - Fluid dynamics
    - Cutting edge Imaging
    - Product development
    - Media/VR simulation

Enterprise DRM "Case Study"

- BioTeam Role - Phase I
  - Work for IS organization
    - objective & vendor agnostic
  - Understand the science
    - Translate to "IT terms"
      - CPU, disk, network, etc.
  - Document scientific requirements and workflows for IS/IT management
  - Propose HPC infrastructure options

Enterprise DRM “Case Study”

- BioTeam Role - Phase II
  - Finalize HPC architecture
  - Assist with RFQ/RFP process
  - Assist with vendor evaluations
  - Manage delivery
  - Install, setup, configure
  - Document
  - Custom onsite training for:
    - System Administrators
    - IS Management
    - Scientific End Users

Enterprise DRM “Case Study”

- Proposed Research Computing Solution:
  - Lots of highly available multi-protocol enterprise storage
    - Some scientists can generate 1TB *per experiment*
    - *Various file types (small vs. large; ascii vs. binary)*
    - *Varied security needs, HPPA/SOX/Audit compliance etc.*
  - Linux compute cluster on Sun hardware
    - Dual-core / dual-processor
  - Large SMP / Large memory servers
  - Virtualized containers for researcher applications
    - Often don’t conform to enterprise security/IT standards …
  - Policy based distributed resource management (DRM)
    - Policies created by Scientific User Board (not IT)

Enterprise DRM “Case Study”

- More about the company:
  - Big, multinational, conservative
  - General rule: no server connectivity to the internet
  - Some IT staff very proud of forcing a 100% Windows research IS environment
  - Little Linux / Open source experience/comfort
  - IT is not a core function of this company
  - Almost all IT functions are immediately outsourced to third party providers

Enterprise DRM “Case Study”

- More about the new research facility:
  - Word from the top:
    - Multi-disciplinary science
    - Massive cross-department collaboration
    - “Open”, “Collaborative”, “Flexible” are the new watchwords
  - What this means:
    - Linux, wiki’s, grids and policy-driven DRM!
    - Plan for multi-site distributed computing
    - Shared infrastructure used by all groups
    - More freedom to be innovative and clever

Researcher DRM Requirements

- Pretty standard really:
  - Policy based resource allocation
  - Array Jobs
  - Job dependencies & resource requests
    - For simple workflows
  - FLEXlm aware
  - Web interface for job submission & monitoring

Management DRM Requirements

- Sorted by priority:
  - Ease of outsourcing
    - Lowest possible administrative burden
    - Quality of support
    - Highest possible resiliency
    - Quality of technical documentation
    - Quality/scope of training
  - Quality/scope of reporting tools
  - “Reasonable” cost
  - Grid buzzword compliant …
  - Play nicely with future WAN-scale computing
  - Software development APIs

Due to these requirements …

I formally recommended the use of Platform LSF

Time out

How businesses purchase DRM

- Key message to impart:
  - *In 2008 ALL* implementations (SGE, PBS, PBSPro, Torque, LSF, etc.) are of very high quality
  - *All* products do “policy based resource allocation on distributed systems” very well
  - Choosing a DRM product now is much harder

How businesses purchase DRM

- Because all DRM's now excel at core scheduling and policy functions …
- Customer's use other factors to choose:
  - Cost
  - Support
  - Training
  - Documentation
  - Additional functionality
    - Embedded or at extra cost

How businesses purchase DRM

- Another key message
  - I am not worried for SGE in any way
    - Excellent system, gaining share all the time
    - Improving at a faster rate than all others

LSF 7.x

Direction, license model & new features

LSF License Model

- Only commercial
  - Complexity comes from layered products
- Must have a license to operate
- Machines are licensed per processor
  - Three classes of LSF server license
  - Multi-core requires additional multi-core license(s)
    - Multi-core licenses only for X86_64/AMD64
      - Other (IBM/Sun/HP) CPUs require a CPU license for each core
- Purchased Licenses:
  - Require use of FLEXlm server
- Demo/evaluation licenses:
  - FLEXlm server not required

LSF License Model

- Server licenses
  - One license per socket/CPU
  - All CPUs on a SMP system must be licensed
  - Licenses are checked out by the Master LIM upon LSF startup
- Client licenses (static)
  - One license per client system
- Client licenses (floating)
  - Any host can consume these entitlements
  - License is held by client until midnight or until the next LSF reconfiguration

LSF License Model

- The three types of LSF Server licenses:
  - B-Class (up to 2 CPUs and 4 GB of memory
  - S-Class (up to 4 CPUs and 16 GB of memory)
  - E-Class (Enterprise, no restriction)
- And
  - Multi-core enablers for x86 and x86_64

LSF License Model

- LSF Server license cost:
  - B-Class: Expensive
  - S-Class: More expensive
  - E-Class: *Wow.*

LSF License Model

- My personal complaints about the license model
  - Knocking me out of “B-Class” because I have a compute node with 4+ GB RAM was simply unacceptable in 2007
    - Extra cost is so significant it can force a recalculation of server/hardware configuration
  - “Cheap” multi-core enablers only available for X86_64/AMD64
    - Significant price penalty for using Power or Sparc architectures

New in LSF 7.0

- An incomplete list of new features …

New in LSF 7.0

- LSF 7.x Stated Performance Goals
  - 5,000 dual-cpu; dual-core hosts
    - Sustain 20 job/query submissions per second
    - Support peak job/query of 100/sec
    - Support 10M completed jobs per day
    - Support 500K active jobs
    - Reconfig and Failover should not exceed 5 minutes
    - Support 8x 192-CPU parallel jobs concurrently

Source:
Personal notes taken at a training class - these figures are not official and could be 100% incorrect. You have been warned.

New in LSF 7.0

- EGO is introduced
- LSF now operates under EGO
  - EGO: “Enterprise Grid Orchestrator”
  - Key concept:
    - EGO is a global “resource broker”
    - LSF is the DRM plug-in for EGO
- Performance Improvements
- Job Application Profiles (!!)

New in LSF 7.0

- Application Profiles
  - Very cool
  - New config file: “lsb.applications”
- What it allows
  - Custom pre/post and starter scripting
  - Custom requeue and exit codes
  - Custom resource requests / limits / rerunnable

New in LSF 7.0

- Application Profiles
- Potential Benefits
  - Custom pre/post/starter scripting is way cool
  - Hide complexity; preconfigure core requirements and settings for common apps
  - Centralized config & control
  - Can potentially greatly reduce the need for dedicated or custom LSF queues

New in 7.0

- I feel the same way about LSF Application Profiles as I do about SGE Resource Quotas
- Both are *very* significant enhancements
- Both likely to have significant positive impact on production user environments

LSF Daemons

LSF Daemons as of 6.0

- LSF 6.x and prior
  - LSF Daemons
    - mbatchd
    - mbschd
    - sbatchd
    - LIM
    - PIM
    - RES

Daemons as of version 7.0

- EGO
  - VEMKD
  - PEM
  - EGOSC
  - LIM
  - PIM
- LSF 7
  - mbatchd
  - mbschd
  - sbatchd
  - RES

Note: EGO becomes the resource broker, LSF becomes the DRM subsystem that consumes EGO resources ...

Daemons on LSF Master Host

- LIM
- PEM
- SBATCHD
- MBATCHD
- MBSCHD
- PIM
- VEMKD
- RES
- EGOSC

Daemons on other hosts

- LIM
- SBATCHD
- PIM
- PEM
- RES

EGO Thoughts (outdated?)

- In 2007 - questionable advantage for customer
- “Placeholder” for future Platform initiatives
- Unofficial word: “Disable EGO in 7.0”
- 2008
  - LSF 7 update 2 out now
  - LSF 7 update 3 next month
  - *Maybe EGO is better*?
- Interesting topic:
  - EGO vs. Sun’s Project Hedeby

Back to LSF & SGE

Remember this slide?

- “Case Study” Mgmt. Needs, priority sorted:
  - Lowest possible administrative burden
  - Support quality
  - Highest possible resiliency
  - Quality of technical documentation
  - Quality/scope of training
  - Quality/scope of reporting tools
  - “Reasonable” cost
  - Grid buzzword compliant …
  - Play nicely with future WAN-scale computing
  - Software development APIs

Web Interface

- Platform has the edge
  - Comes “for free” with LSF
  - Nothing comparable within SGE base
- Tomcat Java application server
  - Web front end for users
    - Submit, monitor, control jobs
  - Web front end for LSF Admins
    - Control hosts and queues

Administrative Burden

- Difficult to quantify …
- My feeling after many years:
  - LSF requires “less work” to operate
    - Installation, Operation, Policies, Troubleshooting
  - Too many SGE best practices are described only on mailing lists
    - *Today’s SGE 6.2 “wiki” announcement could change this*
- Non trivial issue
  - Staff costs far higher than sw license costs

Support

- No clear winner
- Sun & Platform both get good marks
- Not sure if LSF support is 24/7 by default

Support

- Direct quote from LSF customer on the beowulf mailing list:
  - “The last time I reported a bug they had a fix to me inside two hours”
  - “What's more, although we're a big customer now, my experience of them as a small customer in the past, with less than ten nodes, was just as good.”

Resiliency

- LSF failover model is excellent
  - Past issues with reconfig/failover fixed in 7.x
- SGE has good model but still gaps
  - BerkeleyDB issues on NFSv3 filesystems
  - SpoolDB server is a single point of failure
    - Waiting on BDB replication from (Oracle) now …
  - Each choice has tradeoffs:
    - Binary spooling on H/A (non NFSv3) filesystems
    - Classic spooling on H/A filesystem
    - H/A clustering at HW/OS level for qmaster system

Technical Documentation

- LSF has an edge
  - More documentation
  - Slicker presentation/organization/delivery
  - Shorter update frequency
  - Things worth emulating:
    - Custom “Your cluster” getting started doc
    - Quick reference cheatsheet*

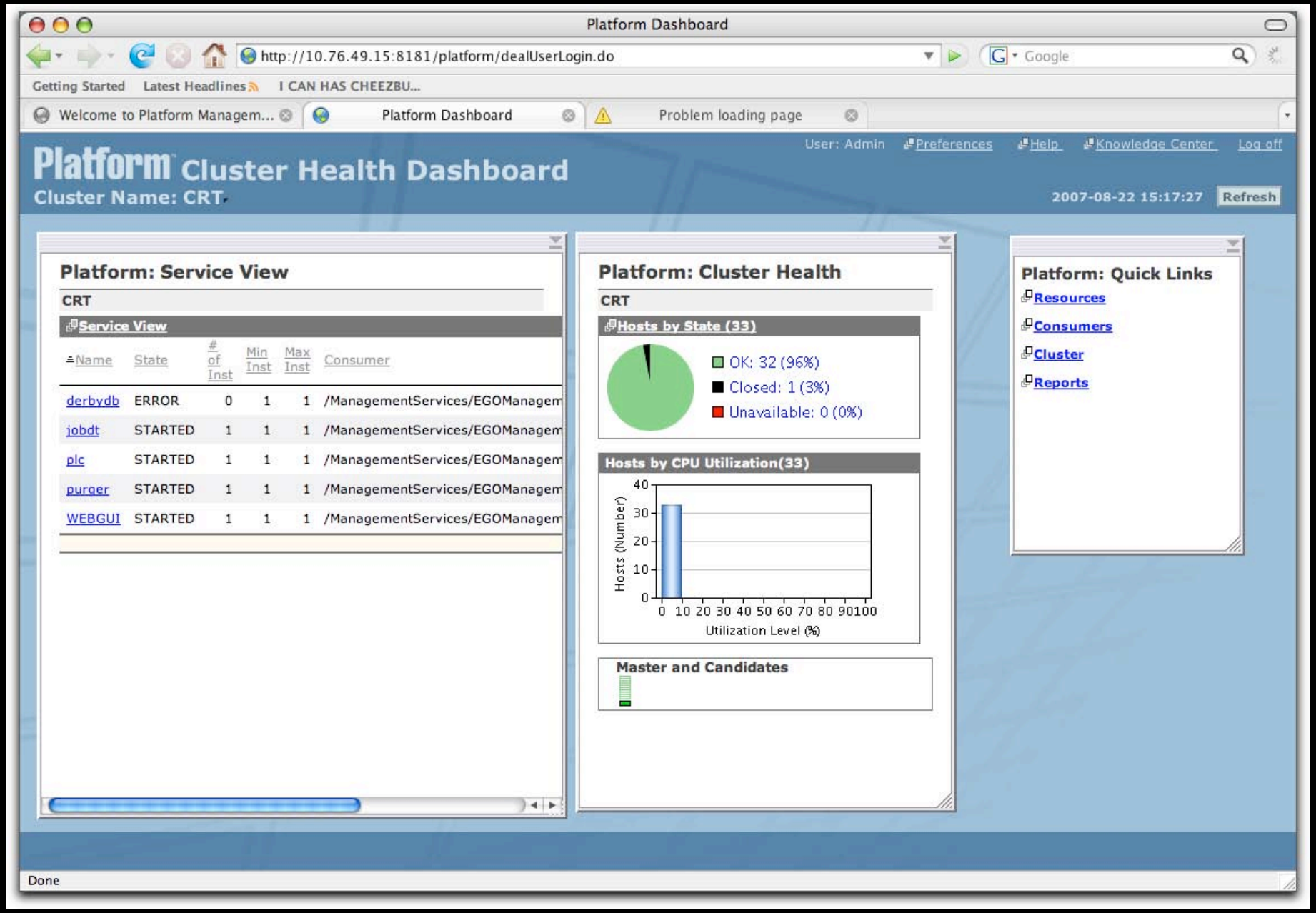
* - http://blog.bioteam.net/2008/02/06/grid-engine-quick-reference-guide/

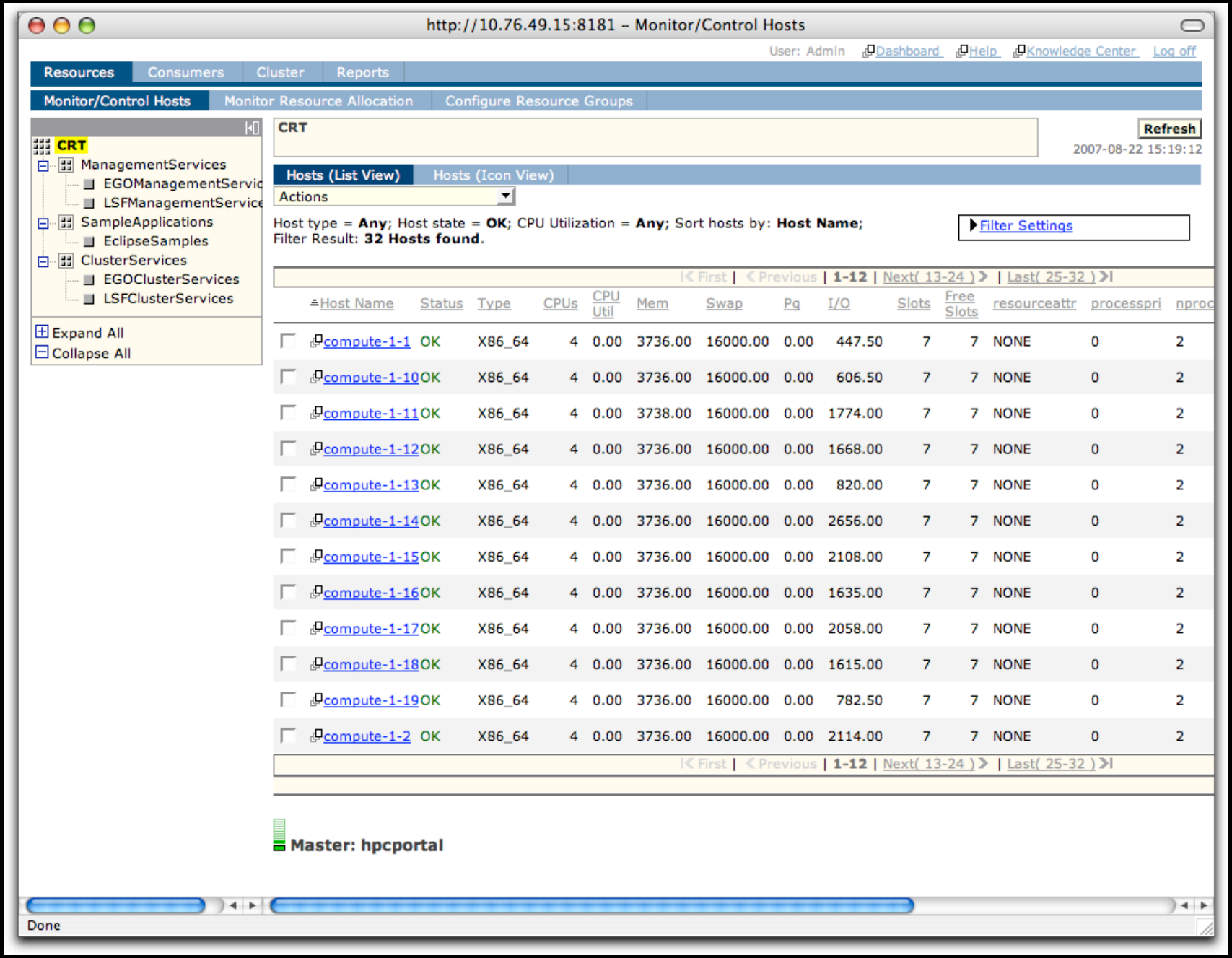
Training

- SunED
  - 1 instructor-led class
  - Teaching 6.0 content
  - Offered twice per year(?)
    - USA only(?)
  - Professional Trainers
    - Trainers have never used SGE in a production setting
  - Caveats
    - Dan Templeton
      - 2x/year @ Georgetown
      - Here!
    - Sun PS and SGE groups can do custom stuff
- Platform
  - Multiple classes
    - Basic, Advanced, Delta
  - Taught semi-monthly
    - All over the world
  - Instructors come from the PS group
    - Many years of hands-on customer facing experience

Reporting Tools

- LSF has the edge
  - “Platform PMC” comes free with base product
  - Platform Analytics layered product for largest enterprises
- ARCo implementation is good, but …
  - Front end has rough edges
  - Users input SQL statements into a web form textarea?


Platform Dashboard
http://10.76.49.15:8181/platform/dealUserLogin.do
Getting Started
Latest Headlines
I CAN HAS CHEEZBU...
Welcome to Platform Managem...
Platform Dashboard
Problem loading page
User: Admin
Preferences
Help
Knowledge Center
Log off
Platform Cluster Health Dashboard
Cluster Name: CRT
2007-08-22 15:17:27
Refresh
Platform: Service View
CRT
Service View
Name
State
# of Inst
Min Inst
Max Inst
Consumer
derbydb ERROR 0 1 1 /ManagementServices/EGOManagem
jobdt STARTED 1 1 1 /ManagementServices/EGOManagem
plc STARTED 1 1 1 /ManagementServices/EGOManagem
purger STARTED 1 1 1 /ManagementServices/EGOManagem
WEBGUI STARTED 1 1 1 /ManagementServices/EGOManagem
Platform: Cluster Health
CRT
Hosts by State (33)
OK: 32 (96%)
Closed: 1 (3%)
Unavailable: 0 (0%)
Hosts by CPU Utilization(33)
Hosts (Number)
0 10 20 30 40
0 10 20 30 40 50 60 70 80 90100
Utilization Level (%)
Master and Candidates
Platform: Quick Links
Resources
Consumers
Cluster
Reports
Done


http://10.76.49.15:8181 – Monitor/Control Hosts
User: Admin
Dashboard
Help
Knowledge Center
Log off
Resources
Consumers
Cluster
Reports
Monitor/Control Hosts
Monitor Resource Allocation
Configure Resource Groups
CRT
ManagementServices
EGOManagementServic
LSFManagementServic
SampleApplications
EclipseSamples
ClusterServices
EGOClusterServices
LSFClusterServices
Expand All
Collapse All
CRT
Refresh
2007-08-22 15:19:12
Hosts (List View)
Hosts (Icon View)
Actions
Host type = Any; Host state = OK; CPU Utilization = Any; Sort hosts by: Host Name;
Filter Result: 32 Hosts found.
Filter Settings
First | Previous | 1-12 | Next(13-24) | Last(25-32)
Host Name Status Type CPUs CPU Util Mem Swap Pg I/O Slots Free Slots resourceattr processpri nproc
compute-1-1 OK X86_64 4 0.00 3736.00 16000.00 0.00 447.50 7 7 NONE 0 2
compute-1-10 OK X86_64 4 0.00 3736.00 16000.00 0.00 606.50 7 7 NONE 0 2
compute-1-11 OK X86_64 4 0.00 3738.00 16000.00 0.00 1774.00 7 7 NONE 0 2
compute-1-12 OK X86_64 4 0.00 3736.00 16000.00 0.00 1668.00 7 7 NONE 0 2
compute-1-13 OK X86_64 4 0.00 3736.00 16000.00 0.00 820.00 7 7 NONE 0 2
compute-1-14 OK X86_64 4 0.00 3736.00 16000.00 0.00 2656.00 7 7 NONE 0 2
compute-1-15 OK X86_64 4 0.00 3736.00 16000.00 0.00 2108.00 7 7 NONE 0 2
compute-1-16 OK X86_64 4 0.00 3736.00 16000.00 0.00 1635.00 7 7 NONE 0 2
compute-1-17 OK X86_64 4 0.00 3736.00 16000.00 0.00 2058.00 7 7 NONE 0 2
compute-1-18 OK X86_64 4 0.00 3736.00 16000.00 0.00 1615.00 7 7 NONE 0 2
compute-1-19 OK X86_64 4 0.00 3736.00 16000.00 0.00 782.50 7 7 NONE 0 2
compute-1-2 OK X86_64 4 0.00 3736.00 16000.00 0.00 2114.00 7 7 NONE 0 2
First | Previous | 1-12 | Next(13-24) | Last(25-32)
Master: hpcportal
Done

Cost

- For many groups and people, LSF will simply be too expensive
  - This is why SGE gains share
    - Most of the market does not need the highest level items
- But …
  - LSF delivers lots of real, usable value over and above base “policy based scheduling on clusters”
  - For groups that require such things, LSF can be cheaper/faster than internal development or extra staff hires

Miscellaneous

- Significant layered products for LSF
  - FLEXlm, Multi-cluster, Interconnects, …
- UnivaUD is stepping up
  - Full cluster stack
    - SGE + Ganglia + ARCo + Globus
    - All fully supported & integrated
- Software APIs
  - LSF exposes full developer APIs
  - DRMAA works but is tightly scoped
    - Future: JMX / JGDI?

End;

- Questions?
- What did I get wrong or miss?

- Questions / Contact
  - Chris Dagdigian
    - Personal: dag@sonsorol.org
    - Corporate: chris@bioteam.net