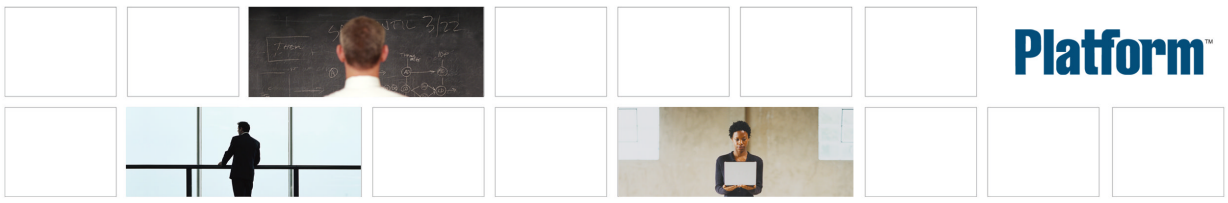
Running Jobs with Platform LSF

Platform LSF™
Version 7.0 Update 5
Release date: March 2009
Last modified: March 16, 2009



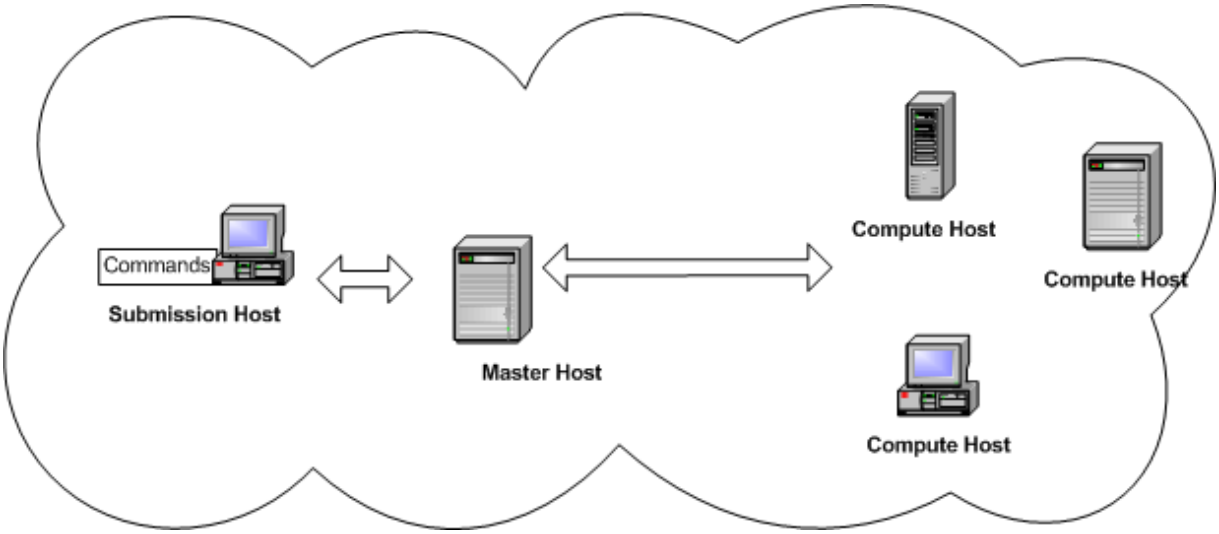
Contents

C H A P T E R

1

About Platform LSF

Clusters, jobs, and queues



Cluster

A group of computers (hosts) running LSF that work together as a single unit, combining computing power and sharing workload and resources. A cluster provides a single-system image for a network of computing resources.

Hosts can be grouped into clusters in a number of ways. A cluster could contain:

- All the hosts in a single administrative group
- All the hosts on one file server or sub-network
- Hosts that perform similar functions

Commands

- `lshosts` — View static resource information about hosts in the cluster
- `bhosts` — View resource and job information about server hosts in the cluster
- `lsid` — View the cluster name
- `lsclusters` — View cluster status and size

Configuration

- Define hosts in your cluster in `lsf.cluster.`*cluster_name*

Tip:

The name of your cluster should be unique. It should not be the same as any host or queue.

Job

A unit of work run in the LSF system. A job is a command submitted to LSF for execution. LSF schedules, controls, and tracks the job according to configured policies.

Jobs can be complex problems, simulation scenarios, extensive calculations, anything that needs compute power.

Commands

- `bjobs` — View jobs in the system
- `bsub` — Submit jobs

Job slot

A job slot is a bucket into which a single unit of work is assigned in the LSF system. Hosts are configured to have a number of job slots available and queues dispatch jobs to fill job slots.

Commands

- `bhosts` — View job slot limits for hosts and host groups
- `bqueues` — View job slot limits for queues
- `busers` — View job slot limits for users and user groups

Configuration

- Define job slot limits in `lsb.resources`.

Job states

LSF jobs have the following states:

- PEND — Waiting in a queue for scheduling and dispatch
- RUN — Dispatched to a host and running
- DONE — Finished normally with zero exit value
- EXITED — Finished with non-zero exit value
- PSUSP — Suspended while pending
- USUSP — Suspended by user
- SSUSP — Suspended by the LSF system
- POST_DONE — Post-processing completed without errors
- POST_ERR — Post-processing completed with errors
- WAIT — Members of a chunk job that are waiting to run

Queue

A clusterwide container for jobs. All jobs wait in queues until they are scheduled and dispatched to hosts.

Queues do not correspond to individual hosts; each queue can use all server hosts in the cluster, or a configured subset of the server hosts.

When you submit a job to a queue, you do not need to specify an execution host. LSF dispatches the job to the best available execution host in the cluster to run that job.

Queues implement different job scheduling and control policies.

Commands

- `bqueues` — View available queues
- `bsub -q` — Submit a job to a specific queue
- `bparams` — View default queues

Configuration

- Define queues in `lsb.queues`

Tip:

The names of your queues should be unique. They should not be the same as the cluster name or any host in the cluster.

First-come, first-served scheduling (FCFS)

The default type of scheduling in LSF. Jobs are considered for dispatch based on their order in the queue.

Hosts

A host is an individual computer in the cluster.

Each host may have more than 1 processor. Multiprocessor hosts are used to run parallel jobs. A multiprocessor host with a single process queue is considered a single machine, while a box full of processors that each have their own process queue is treated as a group of separate machines.

Commands:

- `lsload` — View load on hosts
- `lshosts` — View configuration information about hosts in the cluster including number of CPUS, model, type, and whether the host is a client or server
- `bhosts` — View batch server hosts in the cluster

Tip:

The names of your hosts should be unique. They should not be the same as the cluster name or any queue defined for the cluster.

Submission host

The host where jobs are submitted to the cluster.

Jobs are submitted using the `bsub` command or from an application that uses the LSF API.

Client hosts and server hosts can act as submission hosts.

Commands:

- `bsub` — Submit a job
- `bjobs` — View jobs that are submitted

Execution host

The host where a job runs. Can be the same as the submission host. All execution hosts are server hosts.

Commands

- `bjobs` — View where a job runs

Server host

Hosts that are capable of submitting and executing jobs. A server host runs `sbatchd` to execute server requests and apply local policies.

Commands

- `lshosts` — View hosts that are servers (`server=Yes`)

Configuration

- Server hosts are defined in the `lsf.cluster.`*cluster_name* file by setting the value of `server` to 1

Client host

Hosts that are only capable of submitting jobs to the cluster. Client hosts run LSF commands and act only as submission hosts. Client hosts do not execute jobs or run LSF daemons.

Commands

- `lshosts` — View hosts that are clients (`server=No`)

Configuration

- Client hosts are defined in the `lsf.cluster.`*cluster_name* file by setting the value of `server` to 0

Master host

Where the master LIM and `mbatchd` run. An LSF server host that acts as the overall coordinator for that cluster. Each cluster has one master host to do all job scheduling and dispatch. If the master host goes down, another LSF server in the cluster becomes the master host.

All LSF daemons run on the master host. The LIM on the master host is the master LIM.

Commands

- `lsid` — View the master host name

Configuration

- The master host is the first host listed in the `lsf.cluster.`*cluster_name* file or is defined along with other candidate master hosts by LSF_MASTER_LIST in `lsf.conf`.

LSF daemons

LSF daemon	Role
`mbatchd`	Job requests and dispatch
`mbschd`	Job scheduling
`sbatchd`	Job execution
`res`	

mbatchd

- Master Batch Daemon running on the master host.
- Started by sbatchd.
- Responsible for the overall state of jobs in the system.
- Receives job submission, and information query requests.
- Manages jobs held in queues. Dispatches jobs to hosts as determined by mbschd.

Configuration

- Port number defined in `lsf.conf`.

mbschd

- Master Batch Scheduler Daemon running on the master host.
- Works with mbatchd.
- Started by `mbatchd`.
- Makes scheduling decisions based on job requirements, and policies, and resource availability. Sends scheduling decisions to the mbatchd.

sbatchd

- Slave Batch Daemon running on each server host.
- Receives the request to run the job from mbatchd and manages local execution of the job.
- Responsible for enforcing local policies and maintaining the state of jobs on the host.
- The sbatchd forks a child `sbatchd` for every job. The child sbatchd runs an instance of res to create the execution environment in which the job runs. The child sbatchd exits when the job is complete.

Commands

`badmin hstartup` — Starts sbatchd

`badmin hshutdown` — Shuts down sbatchd

`badmin hrestart` — Restarts sbatchd

Configuration

- Port number defined in `lsf.conf`.

res

- Remote Execution Server (`res`) running on each server host.
- Accepts remote execution requests to provide transparent and secure remote execution of jobs and tasks.

Commands

- `lsadmin resstartup` — Starts `res`
- `lsadmin resshutdown` — Shuts down `res`
- `lsadmin resrestart` — Restarts `res`

Configuration

- Port number defined in `lsf.conf`.

lim

- Load Information Manager (`LIM`) running on each server host.
- Collects host load and configuration information and forwards it to the master LIM running on the master host.
- Reports the information displayed by `lsload` and `lshosts`.
- Static indices are reported when the LIM starts up or when the number of CPUs (ncpus) change. Static indices are:
  - Number of CPUs (ncpus)
  - Number of disks (ndisks)
  - Total available memory (maxmem)
  - Total available swap (maxswp)
  - Total available temp (maxtmp)
  - Dynamic indices for host load collected at regular intervals are:
  - Hosts status (status)
  - 15 second, 1 minute, and 15 minute run queue lengths (r15s, r1m, and r15m)
  - CPU utilization (ut)
  - Paging rate (pg)
  - Number of login sessions (ls)
  - Interactive idle time (it)
  - Available swap space (swp)
  - Available memory (mem)
  - Available temp space (tmp)
  - Disk IO rate (io)

Commands

- `lsadmin limstartup` — Starts LIM
- `lsadmin limshutdown` — Shuts down LIM
- `lsadmin limrestart` — Restarts LIM
- `lsload` — View dynamic load values
- `lshosts` — View static host load values

Configuration

- Port number defined in `lsf.conf`.

Master LIM

- The LIM running on the master host. Receives load information from the LIMs running on hosts in the cluster.
- Forwards load information to mbatchd, which forwards this information to mbschd to support scheduling decisions. If the master LIM becomes unavailable, a LIM on another host automatically takes over.

Commands

- `lsadmin limstartup` — Starts LIM
- `lsadmin limshutdown` — Shuts down LIM
- `lsadmin limrestart` — Restarts LIM
- `lsload` — View dynamic load values
- `lshosts` — View static host load values

Configuration

- Port number defined in `lsf.conf`.

ELIM

External LIM (ELIM) is a site-definable executable that collects and tracks custom dynamic load indices. An ELIM can be a shell script or a compiled binary program, which returns the values of the dynamic resources you define. The ELIM executable must be named elim and located in LSF_SERVERDIR.

pim

- Process Information Manager (PIM) running on each server host.
- Started by LIM, which periodically checks on `pim` and restarts it if it dies.
- Collects information about job processes running on the host such as CPU and memory used by the job, and reports the information to sbatchd.

Commands

- `bjobs` — View job information

Batch jobs and tasks

You can either run jobs through the batch system where jobs are held in queues, or you can interactively run tasks without going through the batch system, such as tests.

Job

A unit of work run in the LSF system. A job is a command submitted to LSF for execution, using the `bsub` command. LSF schedules, controls, and tracks the job according to configured policies.

Jobs can be complex problems, simulation scenarios, extensive calculations, anything that needs compute power.

Commands

- `bjobs` — View jobs in the system
- `bsub` — Submit jobs

Interactive batch job

A batch job that allows you to interact with the application and still take advantage of LSF scheduling policies and fault tolerance. All input and output are through the terminal that you used to type the job submission command.

When you submit an interactive job, a message is displayed while the job is awaiting scheduling. A new job cannot be submitted until the interactive job is completed or terminated.

The `bsub` command stops display of output from the shell until the job completes, and no mail is sent to you by default. Use `Ctrl-C` at any time to terminate the job.

Commands

- `bsub -I` — Submit an interactive job

Interactive task

A command that is not submitted to a batch queue and scheduled by LSF, but is dispatched immediately. LSF locates the resources needed by the task and chooses the best host among the candidate hosts that has the required resources and is lightly loaded. Each command can be a single process, or it can be a group of cooperating processes.

Tasks are run without using the batch processing features of LSF but still with the advantage of resource requirements and selection of the best host to run the task based on load.

Commands

- `lsrun` — Submit an interactive task
- `lsgrun` — Submit an interactive task to a group of hosts

See also LSF utilities such as `ch`, `lsacct`, `lsacctmrg`, `lslogin`, `lsplace`, `lsload`, `lsloadadj`, `lseligible`, `lsmon`, `lstcsh`.

Local task

An application or command that does not make sense to run remotely. For example, the `ls` command on UNIX.

Commands

- `lsltasks` — View and add tasks

Configuration

- `lsf.task` — Configure system-wide resource requirements for tasks
- `lsf.task.cluster` — Configure cluster-wide resource requirements for tasks
- `.lsftasks` — Configure user-specific tasks

Remote task

An application or command that can be run on another machine in the cluster.

Commands

- `lsrtasks` — View and add tasks

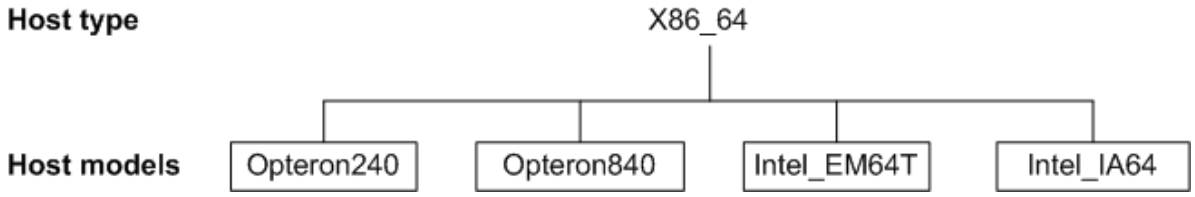
Configuration

- `lsf.task` — Configure system-wide resource requirements for tasks
- `lsf.task.cluster` — Configure cluster-wide resource requirements for tasks
- `.lsftasks` — Configure user-specific tasks

Host types and host models

Hosts in LSF are characterized by host type and host model.

The following example has a host type of X86_64. Host models are Opteron240, Opteron840, Intel_EM64T, Intel_IA64.



Host type

The combination of operating system version and host CPU architecture.

All computers that run the same operating system on the same computer architecture are of the same type — in other words, binary-compatible with each other.

Each host type usually requires a different set of LSF binary files.

Commands

- `lsinfo -t` — View all host types defined in `lsf.shared`

Configuration

- Defined in `lsf.shared`
- Mapped to hosts in `lsf.cluster.`*`cluster_name`*

Host model

The combination of host type and CPU speed (CPU factor) of the computer.

All hosts of the same relative speed are assigned the same host model.

The CPU factor is taken into consideration when jobs are being dispatched.

Commands

- `lsinfo -m` — View a list of currently running models
- `lsinfo -M` — View all models defined in `lsf.shared`

Configuration

- Defined in `lsf.shared`
- Mapped to hosts in `lsf.cluster.`*`cluster_name`*

Users and administrators

LSF user

A user account that has permission to submit jobs to the LSF cluster.

LSF administrator

In general, you must be an LSF administrator to perform operations that will affect other LSF users. Each cluster has one primary LSF administrator, specified during LSF installation. You can also configure additional administrators at the cluster level and at the queue level.

Primary LSF administrator

The first cluster administrator specified during installation and first administrator listed in `lsf.cluster.`*`cluster_name`*. The primary LSF administrator account owns the configuration and log files. The primary LSF administrator has permission to perform clusterwide operations, change configuration files, reconfigure the cluster, and control jobs submitted by all users.

Cluster administrator

May be specified during LSF installation or configured after installation. Cluster administrators can perform administrative operations on all jobs and queues in the cluster. Cluster administrators have the same cluster-wide operational privileges as the primary LSF administrator except that they do not necessarily have permission to change LSF configuration files.

For example, a cluster administrator can create an LSF host group, submit a job to any queue, or terminate another user's job.

Queue administrator

An LSF administrator user account that has administrative permissions limited to a specified queue. For example, an LSF queue administrator can perform administrative operations on the specified queue, or on jobs running in the specified queue, but cannot change LSF configuration or operate on LSF daemons.

Resources

Resource usage

The LSF system uses built-in and configured resources to track resource availability and usage. Jobs are scheduled according to the resources available on individual hosts.

Jobs submitted through the LSF system will have the resources they use monitored while they are running. This information is used to enforce resource limits and load thresholds as well as fairshare scheduling.

LSF collects information such as:

- Total CPU time consumed by all processes in the job
- Total resident memory usage in KB of all currently running processes in a job
- Total virtual memory usage in KB of all currently running processes in a job
- Currently active process group ID in a job
- Currently active processes in a job

On UNIX, job-level resource usage is collected through PIM.

Commands

- `lsinfo` — View the resources available in your cluster
- `bjobs -l` — View current resource usage of a job

Configuration

- SBD_SLEEP_TIME in `lsb.params` — Configures how often resource usage information is sampled by PIM, collected by `sbatchd`, and sent to `mbatchd`

Load indices

Load indices measure the availability of dynamic, non-shared resources on hosts in the cluster. Load indices built into the LIM are updated at fixed time intervals.

Commands

- `lsload -l` — View all load indices
- `bhosts -l` — View load levels on a host

External load indices

Defined and configured by the LSF administrator and collected by an External Load Information Manager (ELIM) program. The ELIM also updates LIM when new values are received.

Commands

- `lsinfo` — View external load indices

Static resources

Built-in resources that represent host information that does not change over time, such as the maximum RAM available to user processes or the number of processors in a machine. Most static resources are determined by the LIM at start-up time.

Static resources can be used to select appropriate hosts for particular jobs based on binary architecture, relative CPU speed, and system configuration.

Load thresholds

Two types of load thresholds can be configured by your LSF administrator to schedule jobs in queues. Each load threshold specifies a load index value:

- loadSched determines the load condition for dispatching pending jobs. If a host's load is beyond any defined `loadSched`, a job will not be started on the host. This threshold is also used as the condition for resuming suspended jobs.
- `loadStop` determines when running jobs should be suspended.

  To schedule a job on a host, the load levels on that host must satisfy both the thresholds configured for that host and the thresholds for the queue from which the job is being dispatched.

  The value of a load index may either increase or decrease with load, depending on the meaning of the specific load index. Therefore, when comparing the host load conditions with the threshold values, you need to use either greater than (>) or less than (<), depending on the load index.

Commands

- `bhosts-l` — View suspending conditions for hosts
- `bqueues -l` — View suspending conditions for queues
- `bjobs -l` — View suspending conditions for a particular job and the scheduling thresholds that control when a job is resumed

Configuration

- `lsb.hosts` — Configure thresholds for hosts
- `lsb.queues` — Configure thresholds for queues

Runtime resource usage limits

Limit the use of resources while a job is running. Jobs that consume more than the specified amount of a resource are signalled or have their priority lowered.

Configuration

- `lsb.queues` — Configure resource usage limits for queues

Hard and soft limits

Resource limits specified at the queue level are hard limits while those specified with job submission are soft limits. See `setrlimit(2)` man page for concepts of hard and soft limits.

Resource allocation limits

Restrict the amount of a given resource that must be available during job scheduling for different classes of jobs to start, and which resource consumers the limits apply to. If all of the resource has been consumed, no more jobs can be started until some of the resource is released.

Configuration

- `lsb.resources` — Configure queue-level resource allocation limits for hosts, users, queues, and projects

Resource requirements (bsub -R)

Restrict which hosts the job can run on. Hosts that match the resource requirements are the candidate hosts. When LSF schedules a job, it collects the load index values of all the candidate hosts and compares them to the scheduling conditions. Jobs are only dispatched to a host if all load values are within the scheduling thresholds.

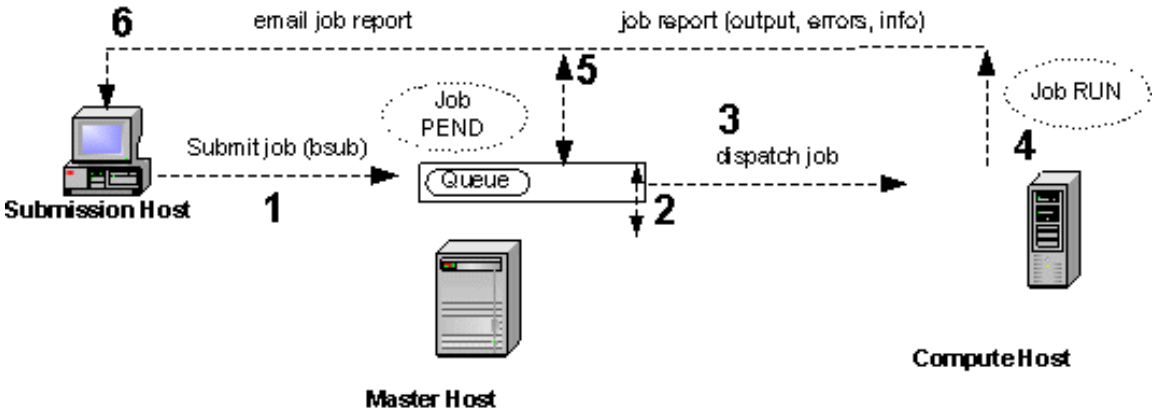
Commands

- `bsub-R` — Specify resource requirement string for a job

Configuration

- `lsb.queues` — Configure resource requirements for queues

Job lifecycle



1 Submit a job

You submit a job from an LSF client or server with the `bsub` command.

If you do not specify a queue when submitting the job, the job is submitted to the default queue.

Jobs are held in a queue waiting to be scheduled and have the PEND state. The job is held in a job file in the `LSF_SHAREDIR/`*cluster_name*`/logdir/info/` directory.

Job ID

LSF assigns each job a unique job ID when you submit the job.

Job name

You can also assign a name to the job with the `-J` option of `bsub`. Unlike the job ID, the job name is not necessarily unique.

2 Schedule job

1. `mbatchd` looks at jobs in the queue and sends the jobs for scheduling to `mbschd` at a preset time interval (defined by the parameter JOB_SCHEDULING_INTERVAL in `lsb.params`).
2. `mbschd` evaluates jobs and makes scheduling decisions based on:
   - Job priority
   - Scheduling policies
   - Available resources
3. `mbschd` selects the best hosts where the job can run and sends its decisions back to `mbatchd`.

Resource information is collected at preset time intervals by the master LIM from LIMs on server hosts. The master LIM communicates this information to `mbatchd`, which in turn communicates it to `mbschd` to support scheduling decisions.

3 Dispatch job

As soon as `mbatchd` receives scheduling decisions, it immediately dispatches the jobs to hosts.

4 Run job

`sbatchd` handles job execution. It:

1. Receives the request from `mbatchd`
2. Creates a child `sbatchd` for the job
3. Creates the execution environment
4. Starts the job using `res`

The execution environment is copied from the submission host to the execution host and includes the following:

- Environment variables needed by the job
- Working directory where the job begins running
- Other system-dependent environment settings, for example:
  - On UNIX and Linux, resource limits and `umask`
  - On Windows, desktop and Windows root directory

  The job runs under the user account that submitted the job and has the status RUN.

5 Return output

When a job is completed, it is assigned the DONE status if the job was completed without any problems. The job is assigned the EXIT status if errors prevented the job from completing.

`sbatchd` communicates job information including errors and output to `mbatchd`.

6 Send email to client

`mbatchd` returns the job output, job error, and job information to the submission host through email. Use the `-o` and `-e` options of `bsub` to send job output and errors to a file.

Job report

A job report is sent by email to the LSF client and includes:

- Job information such as:
  - CPU use
  - Memory use
  - Name of the account that submitted the job
- Job output
- Errors

CHAPTER

2

Working with Jobs

Submitting jobs (bsub)

Submit a job

To have your work run on the cluster, you must submit jobs. You have many options when submitting a job.

The following is the most basic way of submitting a job.

1. To submit a job, run `bsub`.

   If you do not specify any options, the job is submitted to the default queue configured by the LSF administrator (usually queue normal)

   For example, if you submit the job `my_job` without specifying a queue, the job goes to the default queue.

   ```
   bsub my_job
   Job <1234> is submitted to default queue <normal>
   ```

   In the above example, 1234 is the job ID assigned to this job, and `normal` is the name of the default job queue.

Your job remains pending until all conditions for its execution are met. Each queue has execution conditions that apply to all jobs in the queue, and you can specify additional conditions when you submit the job

When you submit jobs, you can also specify an execution host or a range of hosts, a queue, and start and termination times, as well as a wide range of other job options.

About submitting a job to a specific queue

Job queues represent different job scheduling and control policies. All jobs submitted to the same queue share the same scheduling and control policy. Each job queue can use a configured subset of server hosts in the cluster; the default is to use all server hosts.

System administrators can configure job queues to control resource access by different users and types of application. Users select the job queue that best fits each job.

The default queue is normally suitable to run most jobs, but the default queue may assign your jobs a very low priority, or restrict execution conditions to minimize interference with other jobs. If automatic queue selection is not satisfactory, choose the most suitable queue for each job.

The factors affecting which queue to choose are user access restrictions, size of job, resource limits of the queue, scheduling priority of the queue, active time windows of the queue, hosts used by the queue, scheduling load conditions, and the queue description displayed by the `bqueues -l` command.

View available queues

You can submit jobs to a queue as long as its STATUS is Open. However, jobs are not dispatched unless the queue is Active.

- To see available queues, run `bqueues`.
- To display only the queues that accept jobs from specific users or user groups, run `bqueues -u user_name`.

- To display only the queues that use specific host name or host group name to run jobs, run `bqueues -m host_name`.

Submit a job to a queue

The following examples are based on the queues defined in the default configuration. Your LSF administrator may have configured different queues.

- To run a job during off hours because the job generates very high load to both the file server and the network, you can submit it to the night queue:

  ```
  bsub -q "night" my_job
  ```
- If you have an urgent job to run, you may want to submit it to the priority queue:

  ```
  bsub -q "priority" my_job
  ```
- If you want to use hosts owned by others and you do not want to bother the owners, you may want to run your low priority jobs on the idle queue so that as soon as the owner comes back, your jobs get suspended:

  ```
  bsub -q "idle" my_job
  ```
- If you are running small jobs and do not want to wait too long to get the results, you can submit jobs to the short queue to be dispatched with higher priority:

  ```
  bsub -q "short" my_job
  ```

  Tip:

  Make sure your jobs are short enough that they are not killed for exceeding the CPU time limit of the queue (check the resource limits of the queue, if any).
- If your job requires a specific execution environment, you may need to submit it to a queue that has a particular job starter defined. LSF administrators are able to specify a queue-level job starter as part of the queue definition; ask them for the name of the queue and configuration details.

Submit a job associated with a project (bsub -P)

Project names are logged in `lsb.acct`. You can use the `bacct` command to gather accounting information on a per-project basis.

1. To associate a project name with a job, run `bsub -P` *project_name*.

   On systems running IRIX 6, before the submitted job begins execution, a new array session is created and the project ID corresponding to the project name is assigned to the session.

Submit a job associated with a user group (bsub -G)

Note:

This task is only useful with fairshare scheduling.

You can specify any user group to which you belong as long as it does not contain any subgroups. You must be a direct member of the specified user group.

User groups in non-leaf nodes cannot be specified because it will cause ambiguity in determining the correct shares given to a user.

1. To submit a job and associate it with a specified user group, run `bsub -G` *user_group*.

For example, to submit the job `myjob` associated to user group `special`:

```
bsub -G special myjob
```

Submit a job with a job name (bsub -J)

You can associate a name with a job or job array.

Job names can contain up to 4094 characters for UNIX and Linux, or up to 255 characters for Windows.

You can later use the job name to identify the job. The job name need not be unique.

1. Use `bsub -J` *job_name* to submit a job and assign a job name to it.

   For example, to submit a job and assign the name `my_job`:

   ```
   bsub -J my_job sleep 1000
   ```

Submit a job to a service class (bsub -sla)

You submit jobs to a service class as you would to a queue, except that a service class is a higher level scheduling policy that makes use of other, lower level ProductNameAbbreviation policies like queues and host partitions to satisfy the service-level goal that the service class expresses.

The service class name where the job is to run is configured in `lsb.serviceclasses`. If the SLA does not exist or the user is not a member of the service class, the job is rejected.

Outside of the configured time windows, the SLA is not active, and ProductNameAbbreviation schedules jobs without enforcing any service-level goals. Jobs will flow through queues following queue priorities even if they are submitted with `-sla`.

Tip:

You should submit your jobs with a run time limit (`-W` option) or the queue should specify a run time limit (RUNLIMIT in the queue definition in `lsb.queues`). If you do not specify a run time limit, ProductNameAbbreviation automatically adjusts the optimum number of running jobs according to the observed run time of finished jobs.

1. To submit a job to a service class for SLA-driven scheduling, run `bsub -sla` *service_class_name*.

   For example:

   ```
   bsub -W 15 -sla Kyuquot sleep 100
   ```

   submits the UNIX command `sleep` together with its argument 100 as a job to the service class named `Kyuquot`.

Submit a job under a job group (bsub -g)

The job group does not have to exist before submitting the job.

1. To submit a job into a job group, run `bsub -g` . For example:

   ```
   bsub -g /risk_group/portfolio1/current myjob
   Job <105> is submitted to default queue.
   ```

   Submits `myjob` to the job group `/risk_group/portfolio1/current`.

If group `/risk_group/portfolio1/current` exists, job 105 is attached to the job group.

If group `/risk_group/portfolio1/current` does not exist, ProductNameAbbreviation checks its parent recursively, and if no groups in the hierarchy exist, all three job groups are created with the specified hierarchy and the job is attached to group.

Modify pending jobs (bmod)

If your submitted jobs are pending (bjobs shows the job in PEND state), you can modify job submission parameters. You can also modify entire job arrays or individual elements of a job array.

- To replace the job command line, run `bmod -Z "new_command"`.

  For example: bmod -Z "myjob file" 101.
- To change a specific job parameter, run `bmod -b`.

  The specified options replace the submitted options.

  The following example changes the start time of job 101 to 2:00 a.m.: bmod -b 2:00 101
- To reset an option to its default submitted value (undo a bmod), append the n character to the option name and do not include an option value.

  The following example resets the start time for job 101 back to its default value: `bmod -bn 101`

Modify the service class of a job

You can attach or detach a service class of a job after the job has been submitted.

Restriction:

You cannot:

- Use -sla with other bmod options
- Move job array elements from one service class to another, only entire job arrays
- Modify the service class of jobs already attached to a job group

- Use the -sla option of bmod to modify the service class a job is attached to, or to attach a submitted job to a service class.

  ```
  bmod -sla Kyuquot 2307
  ```

  Attaches job 2307 to the service class Kyuquot.
- Use bmod -slan to detach a job from a service class.

  For example:

  ```
  bmod -slan 2307
  ```

  Detaches job 2307 from the service class Kyuquot.

Modify a job submitted to a job group

You can modify your own job groups and job groups that other users create under your job groups. The LSF administrator can modify job groups of all users.

Restriction:

The command `bmod -g` cannot be combined with other `bmod` options. It can only operate on pending jobs.

It cannot operate on running or finished jobs.

You cannot move job array elements from one job group to another, only entire job arrays. A job array can only belong to one job group at a time.

You cannot modify the job group of a job attached to a service class.

- To specify a job group path to move a job or a job array from one job group to another, run `bmod -g`.

  For example:

  ```
  bmod -g /risk_group/portfolio2/monthly 105
  ```

  Moves job 105 to job group /risk_group/portfolio2/monthly.

  Like `bsub -g`, if the job group does not exist, LSF creates it.

Modify running jobs

Once you submit a job and it is running, you can modify some of the job options, including resource reservation, CPU limit, memory limit, and others.

Restriction:

You cannot modify remote running jobs in a MultiCluster environment.

- Modify resource reservation
- Modify other selected job options

Modify resource reservation

A job is usually submitted with a resource reservation for the maximum amount required. Use this command to decrease the reservation, allowing other jobs access to the resource.

Note:

You can modify additional job options by setting LSB_MOD_ALL_JOBS in `lsf.conf`.

1. Run `bmod -R` to modify the resource reservation for a running job.

   For example, to set the resource reservation for job 101 to 25MB of memory and 50 MB of swap space:

   ```
   bmod -R "rusage[mem=25:swp=50]" 101
   ```

Modify job options

LSB_MOD_ALL_JOBS is specified in `lsf.conf`.

To modify the CPU limit of running jobs, LSB_JOB_CPULIMIT=Y must be defined in `lsf.conf`.

To modify the memory limit of running jobs, LSB_JOB_MEMLIMIT=Y must be defined in `lsf.conf`.

To modify the name of job error file for a running job, you must use `bsub -e` or `bmod -e` to specify an error file before the job starts running.

If LSB_MOD_ALL_JOBS is specified in `lsf.conf`, the job owner or the LSF administrator can use the `bmod` command to modify some job options for running jobs.

The modified resource limits cannot exceed the resource limits defined in the queue.

Restriction:

You cannot use these options in combination with other `bmod` options. An error message is issued and the modification fails if these options are used on running jobs in combination with other `bmod` options.

- CPU limit (`-c` [*hour*:]*minute*[/*host_name* | /*host_model*] | `-cn`)
- Memory limit (`-M` *mem_limit* | `-Mn`)

- Rerunnable jobs (`-r` | `-rn`)
- Standard error file name (`-e` *error_file* | `-en`)
- Standard output file name (`-o` *output_file* | `-on`)
- Run limit (`-W` *run_limit*[/*host_name* | /*host_model*] | `-Wn`)

About controlling jobs

LSF controls jobs dispatched to a host to enforce scheduling policies or in response to user requests.

The LSF system performs the following actions on a job:

- Suspend by sending a SIGSTOP signal
- Resume by sending a SIGCONT signal
- Terminate by sending a SIGKILL signal

On Windows, equivalent functions have been implemented to perform the same tasks.

Kill a job (bkill)

You can cancel a job from running or pending by killing it.

If your job is part of a job group, the command `bkill` only kills jobs in the job group you specify. It does not kill jobs in lower level job groups in the path.

1. To kill job 3421:

   ```
   bkill 3421
   ```

   a) Use the `-g` option of `bkill` and specify a job group path to terminate jobs in a job group. For example:

   ```
   bkill -g /risk_group 106
   ```

   b) Use job ID 0 (zero) to terminate all jobs in a job group:

   ```
   bkill -g /risk_group 0
   ```

Forcing removal of a job from LSF

If a job cannot be killed in the operating system, you can force the removal of the job from LSF.

The `bkill -r` command removes a job from the system without waiting for the job to terminate in the operating system. This sends the same series of signals as `bkill` without `-r`, except that the job is removed from the system immediately, the job is marked as EXIT, and job resources that LSF monitors are released as soon as LSF receives the first signal.

1. Use `bkill -r` to force the removal of the job from LSF.

About suspending and resuming jobs (bstop and bresume)

You can resume or suspend a job using the `bstop` and `bresume` commands.

A job can be suspended by its owner or the LSF administrator with the `bstop` command. These jobs are considered user-suspended and are displayed by `bjobs` as USUSP.

When the user restarts the job with the `bresume` command, the job is not started immediately to prevent overloading. Instead, the job is changed from USUSP to SSUSP (suspended by the system). The SSUSP job is resumed when host load levels are within the scheduling thresholds for that job, similarly to jobs suspended due to high load.

If a user suspends a high priority job from a non-preemptive queue, the load may become low enough for LSF to start a lower priority job in its place. The load created by the low priority job can prevent the high priority job from resuming. This can be avoided by configuring preemptive queues.

The command `bstop` sends the following signals to the job:

- SIGTSTP for parallel or interactive jobs

  SIGTSTP is caught by the master process and passed to all the slave processes running on other hosts.
- SIGSTOP for sequential jobs

  SIGSTOP cannot be caught by user programs. The SIGSTOP signal can be configured with the LSB_SIGSTOP parameter in `lsf.conf`.

Allow users to resume jobs

If ENABLE_USER_RESUME=Y in `lsb.params`, you can resume your own jobs that have been suspended by the administrator.

Suspend a job

You can suspend or stop a job that is already running.

You must be an administrator or the user who submitted the job.

When you stop a job, it is suspended until you choose to resume it.

Suspending a job causes your job to go into USUSP state if the job is already started, or to go into PSUSP state if your job is pending.

By default, jobs that are suspended by the administrator can only be resumed by the administrator or `root`; users do not have permission to resume a job suspended by another user or the administrator. Administrators can resume jobs suspended by users or administrators. Administrators can also enable users to resume their own jobs that have been stopped by an administrator.

1. Use the `-g` option of `bstop` and specify a job group path to suspend jobs in a job group

   ```
   bstop -g /risk_group 106
   Job <106> is being stopped
   ```

   Use job ID 0 (zero) to suspend all jobs in a job group:

   ```
   bstop -g /risk_group/consolidate 0
   Job <107> is being stopped
   Job <108> is being stopped
   Job <109> is being stopped
   ```

Resume a job

You can resume a job that was suspended using `bstop`.

You must be the same user who suspended the job. If your job was suspended by the administrator, you cannot resume it; the administrator or `root` must resume the job for you.

Resuming a user-suspended job does not put your job into RUN state immediately. If your job was running before the suspension, `bresume` first puts your job into SSUSP state and then waits for `sbatchd` to schedule it according to the load conditions.

1. From the command line, run `bresume` and specify a job group path to resume suspended jobs in a job group.

   For example, to resume job 3421, enter `bresume 3421`.

Move a job to the bottom of a queue (bbot)

Use `bbot` to move jobs relative to your last job in the queue.

You must be an LSF administrator or the user who submitted the job.

By default, LSF dispatches jobs in a queue in the order of arrival (that is, first-come-first-served), subject to availability of suitable server hosts.

Use the `bbot` command to change the position of pending jobs, or of pending job array elements, to affect the order in which jobs are considered for dispatch. Users can only change the relative position of their own jobs, and LSF administrators can change the position of any users' jobs.

- To move a job to the bottom of the queue, run `bbot`.

  For example, `bbot 5311`.

If invoked by a regular user, `bbot` moves the selected job after the last job with the same priority submitted by the user to the queue.

If invoked by the LSF administrator, `bbot` moves the selected job after the last job with the same priority submitted to the queue.

Move a job to the top of a queue (btop)

Use `btop` to move jobs relative to your first job in the queue.

By default, LSF dispatches jobs in a queue in the order of arrival (that is, first-come-first-served), subject to availability of suitable server hosts.

Use the `btop` command to change the position of pending jobs, or of pending job array elements, to affect the order in which jobs are considered for dispatch. Users can only change the relative position of their own jobs, and LSF administrators can change the position of any users' jobs.

- To move a job to the top of the queue, run `btop`.

  In the following example, job 5311 is moved to the top of the queue. Since job 5308 is already running, job 5311 is placed in the queue after job 5308.

  Tip:

  Note that `user1`'s job is still in the same position on the queue. `user2` cannot use `btop` to get extra jobs at the top of the queue; when one of his jobs moves up the queue, the rest of his jobs move down.

```
bjobs -u all
JOBID USER  STAT  QUEUE    FROM_HOST  EXEC_HOST  JOB_NAME   SUBMIT_TIME
5308  user2 RUN   normal   hostA      hostD      /s500      Oct 23 10:16
5309  user2 PEND  night    hostA                 /s200      Oct 23 11:04
5310  user1 PEND  night    hostB                 /myjob     Oct 23 13:45
5311  user2 PEND  night    hostA                 /s700      Oct 23 18:17
btop 5311
Job <5311> has been moved to position 1 from top.
bjobs -u all
JOBID USER  STAT  QUEUE    FROM_HOST  EXEC_HOST  JOB_NAME   SUBMIT_TIME
5308  user2 RUN   normal   hostA      hostD      /s500      Oct 23 10:16
5311  user2 PEND  night    hostA                 /s200      Oct 23 18:17
5310  user1 PEND  night    hostB                 /myjob     Oct 23 13:45
5309  user2 PEND  night    hostA                 /s700      Oct 23 11:04
```

If invoked by a regular user, `btop` moves the selected job before the first job with the same priority submitted by the user to the queue.

If invoked by the LSF administrator, `btop` moves the selected job before the first job with the same priority submitted to the queue.

Control jobs in job groups

Stop jobs (bstop)

1. To suspend jobs in a job group, run `bstop -g` and specify a job group path.

   ```
   bstop -g /risk_group 106
   ```

   Job <106> is being stopped
2. To suspend all jobs in a job group, use job ID 0 (zero).

   ```
   bstop -g /risk_group/consolidate 0
   Job <107> is being stopped
   Job <108> is being stopped
   Job <109> is being stopped
   ```

Resume jobs (bresume)

1. To resume suspended jobs in a job group, run bresume -g and specify a job group path

   ```
   bresume -g /risk_group 106
   ```

   Job <106> is being resumed
2. To resume all jobs in a job group, use job ID 0 (zero).

   ```
   bresume -g /risk_group 0
   Job <109> is being resumed
   Job <110> is being resumed
   Job <112> is being resumed
   ```

Terminate jobs (bkill)

The command `bkill` only kills jobs in the job group you specify. It does not kill jobs in lower level job groups in the path.

1. To terminate jobs in a job group, run `bkill -g` and specify a job group path.

   ```
   bkill -g /risk_group 106
   ```

   Job <106> is being terminated
2. To terminate all jobs in a job group, run job ID 0 (zero).

   ```
   bkill -g /risk_group 0
   Job <1413> is being terminated
   Job <1414> is being terminated
   Job <1415> is being terminated
   Job <1416> is being terminated
   ```

Delete job groups (bgdel)

The job group cannot contain any jobs.

1. To remove a job group, run bgdel.

```
bgdel /risk_group
```

```
Job group /risk_group is deleted.
```

The job group /risk_group and all its subgroups are deleted.

Submit a job to specific hosts

Specify which hosts a job can run on.

When several hosts can satisfy the resource requirements of a job, the hosts are ordered by load. However, in certain situations it may be desirable to override this behavior to give preference to specific hosts, even if they are more heavily loaded.

Note:

By specifying a single host, you can force your job to wait until that host is available and then run on that host.

1. To indicate that a job must run on one of the specified hosts, use the `bsub -m "hostA hostB ..."` option.

   For example:

```
bsub -m "hostA hostD hostB" myjob
```

Submit a job with specific resources

Configure a new resource that your job runs with.

You must be the LSF administrator.

If you have applications that need specific resources, you should create a new Boolean resource and configure that resource for the appropriate hosts in the cluster.

If you specify a host list using the `-m` option of `bsub`, you must change the host list every time you add a new host that supports the desired resources. By using a Boolean resource, the LSF administrator can add, move, or remove resources without forcing users to learn about changes to resource configuration.

Queues and host preference

A queue can define host preferences for jobs. Host preferences specified by `bsub -m` override the queue specification.

In the queue definition in `lsb.queues`, use the HOSTS parameter to list the hosts or host groups to which the queue can dispatch jobs.

Use the not operator (~) to exclude hosts or host groups from the list of hosts to which the queue can dispatch jobs. This is useful if you have a large cluster, but only want to exclude a few hosts from the queue definition.

Specify different levels of host preference

Submit a job with varying levels of host preferences.

You can indicate different levels of preference by specifying a number after the plus sign (`+`). The larger the number, the higher the preference for that host or host group. You can also specify the + with the keyword `others`.

1. Submit a job indicating a host preference.

For example,

```
bsub -m "groupA+2 groupB+1 groupC" myjob
```

In this example, LSF gives first preference to hosts in `groupA`, second preference to hosts in `groupB` and last preference to those in `groupC`. Ordering within a group is still determined by load.

Submit a job with resource requirements

Submit a job that requires specific resources to run.

1. Submit a job indicating a resource requirement.

   For example, to submit a job which will run on Solaris 7 or Solaris 8:

   ```
   bsub -R "sol7 || sol8" myjob
   ```

   When you submit a job, you can also exclude a host by specifying a resource requirement using `hname` resource:

   ```
   bsub -R "hname!=hostb && type==sgi6" myjob
   ```

Using LSF with non-shared file space

LSF is usually used in networks with shared file space. When shared file space is not available, use the `bsub -f` command to have LSF copy needed files to the execution host before running the job, and copy result files back to the submission host after the job completes.

LSF attempts to run the job in the directory where the `bsub` command was invoked. If the execution directory is under the user's home directory, `sbatchd` looks for the path relative to the user's home directory. This handles some common configurations, such as cross-mounting user home directories with the `/net` automount option.

If the directory is not available on the execution host, the job is run in `/tmp`. Any files created by the batch job, including the standard output and error files created by the `-o` and `-e` options to `bsub`, are left on the execution host.

LSF provides support for moving user data from the submission host to the execution host before executing a batch job, and from the execution host back to the submitting host after the job completes. The file operations are specified with the `-f` option to `bsub`.

LSF uses the `lsrcp` command to transfer files. `lsrcp` contacts RES on the remote host to perform file transfer. If RES is not available, the UNIX `rcp` command is used.

bsub -f

The `-f "[`*local_file operator* `[`*remote_file*`]]"` option to the `bsub` command copies a file between the submission host and the execution host. To specify multiple files, repeat the `-f` option.

local_file

File name on the submission host.

remote_file

File name on the execution host.

The files *local_file* and *remote_file* can be absolute or relative file path names. You must specify at least one file name. When the file *remote_file* is not specified, it is assumed to be the same as *local_file*. Including *local_file* without the operator results in a syntax error.

operator

Operation to perform on the file. The operator must be surrounded by white space.

Valid values for *operator* are:

- >

  local_file on the submission host is copied to *remote_file* on the execution host before job execution. *remote_file* is overwritten if it exists.
- <

  remote_file on the execution host is copied to *local_file* on the submission host after the job completes. *local_file* is overwritten if it exists.
- <<

  remote_file is appended to *local_file* after the job completes. *local_file* is created if it does not exist.
- ><, <>

  Equivalent to performing the > and then the < operation. The file *local_file* is copied to *remote_file* before the job executes, and *remote_file* is copied back, overwriting *local_file*, after the job completes. <> is the same as ><

If the submission and execution hosts have different directory structures, you must ensure that the directory where *remote_file* and *local_file* will be placed exists. ProductNameAbbreviation tries to change the directory to the same path name as the directory where the `bsub` command was run. If this directory does not exist, the job is run in your home directory on the execution host.

You should specify *remote_file* as a file name with no path when running in non-shared file systems; this places the file in the job's current working directory on the execution host. This way the job will work correctly even if the directory where the `bsub` command is run does not exist on the execution host. Be careful not to overwrite an existing file in your home directory.

About resource reservation

When a job is dispatched, the system assumes that the resources that the job consumes will be reflected in the load information. However, many jobs do not consume the resources they require when they first start. Instead, they will typically use the resources over a period of time.

For example, a job requiring 100 MB of swap is dispatched to a host having 150 MB of available swap. The job starts off initially allocating 5 MB and gradually increases the amount consumed to 100 MB over a period of 30 minutes. During this period, another job requiring more than 50 MB of swap should not be started on the same host to avoid over-committing the resource.

You can reserve resources to prevent overcommitment by LSF. Resource reservation requirements can be specified as part of the resource requirements when submitting a job, or can be configured into the queue level resource requirements.

How resource reservation works

When deciding whether to schedule a job on a host, LSF considers the reserved resources of jobs that have previously started on that host. For each load index, the amount reserved by all jobs on that host is summed up and subtracted (or added if the index is increasing) from the current value of the resources as reported by the LIM to get amount available for scheduling new jobs:

```
available amount = current value - reserved amount for all jobs
```

View resource information

You can view host - or queue-level resource information.

1. View host-level resource information.
   a) To view the amount of resources reserved on each host, run `bhosts -l`.
   b) To view information about shared resources, run `bhosts -s`.
2. View the queue-level resource information.
   a) To see the resource usage configured at the queue level, run `bqueues -l`.

Submit a job with resource requirements

1. To specify resource reservation at the job level, use `bsub -R` and include the resource usage section in the resource requirement (rusage) string.

   For example:

   ```
   bsub -R "rusage[tmp=30:duration=30:decay=1]" myjob
   ```

   LSF reserves 30 MB of temp space for the job. As the job runs, the amount reserved will decrease at approximately 1 MB/minute such that the reserved amount is 0 after 30 minutes.

Submit a job with start or termination times

You can specify a time of day at which to start or stop a job.

By default, LSF dispatches jobs as soon as possible, and then allows them to finish, although resource limits might terminate the job before it finishes.

If you do not want to start your job immediately when you submit it, specify a start time for LSF to start it.

You can also specify a time after which the job should be terminated.

1. Submit a job with a start time.

   ```
   bsub -b 5:00 myjob
   ```

   This example submits a job that remains pending until after the local time on the master host reaches 5 a.m.
2. Submit a job with a termination time.

   ```
   bsub -t 11:12:20:30 myjob
   ```

   The job called `myjob` is submitted to the default queue. If the job is still running on November 12 at 8:30 p.m., it will be killed.

Submit a job with resource requirements

Submit a job that requires specific resources to run.

1. Submit a job indicating a resource requirement.

   For example, to submit a job which will run on Solaris 7 or Solaris 8:

   ```
   bsub -R "sol7 || sol8" myjob
   ```

   When you submit a job, you can also exclude a host by specifying a resource requirement using `hname` resource:

   ```
   bsub -R "hname!=hostb && type==sgi6" myjob
   ```

Submit a job with compute unit resource requirements

Submit a job with compute unit resource requirements.

1. Submit a job to run on a certain number of compute units or fewer using the keyword maxcus in the resource requirement cu string.

   To submit a job which will run over 4 or fewer compute units of type rack for example:

   ```
   bsub -R "cu[type=rack:maxcus=4]" my_job
   ```
2. Submit a job to run balanced over the fewest possible compute units using the keyword balance in the resource requirement cu string.

   To submit a job which will run evenly distributed over the fewest compute units and use a total of 64 slots, for example:

   ```
   bsub -n 64 -R "cu[balance]" my_job
   ```

   Possible resource allocations (in order of preference) are:

   - 64 slots on 1 enclosure
   - 32 slots on 2 enclosures
   - 22 slots on 1 enclosure and 21 slots on 2 enclosures
   - 16 slots on 4 enclosures
3. Submit a job to run on compute units few or many available slots use the keyword pref.

   To submit a job to run on compute units with the fewest slots for example:

   ```
   bsub -R "cu[pref=minavail]" my_job
   ```

CHAPTER

3

Monitoring Jobs

View information about jobs

1. Use the `bjobs` and `bhist` commands to view information about jobs:
   - `bjobs` reports the status of jobs and the various options allow you to display specific information.
   - `bhist` reports the history of one or more jobs in the system.

   You can also find jobs on specific queues or hosts, find jobs submitted by specific projects, and check the status of specific jobs using their job IDs or names.

View unfinished jobs

1. Run `bjobs` to view the status of LSF jobs.

   When no options are specified, `bjobs` displays information about jobs in the PEND, RUN, USUSP, PSUSP, and SSUSP states for the current user.

View all jobs

You can display information about jobs that are both running and those recently finished (PEND, RUN, USUSP, PSUSP, SSUSP, DONE, and EXIT statuses)

1. Run `bjobs -a` .

   All your jobs that are still in the system and jobs that have recently finished are displayed.

View running jobs

You can display information about only jobs that are running (RUN status).

1. Run `bjobs -r`.

   All your running jobs are displayed.

View pending reasons for jobs

When you submit a job, it may be held in the queue before it starts running and it may be suspended while running. You can find out why jobs are pending or in suspension with the `bjobs -p` option.

1. Run `bjobs -p`.

   Displays information for pending jobs (PEND state) and their reasons. There can be more than one reason why the job is pending.

   The pending reasons alsodisplay the number of hosts for each condition.
2. To get specific host names along with pending reasons, `run bjobs -lp`.
3. To view the pending reasons for all users, run `bjobs -p -u all`.

View suspending reasons for jobs

When you submit a job, it may be held in the queue before it starts running and it may be suspended while running.

1. Run `bjobs -s`.

Displays information for suspended jobs (SUSP state) and their reasons. There can be more than one reason why the job is suspended.

The pending reasons alsodisplay the number of hosts for each condition.

2. To get specific host names along with suspend reasons, `run bjobs -ls`.
3. To view the suspend reasons for all users, run `bjobs -s -u all`.

View detailed job information

The `-l` option of `bjobs` displays detailed information about job status and parameters, such as the job's current working directory, parameters specified when the job was submitted, and the time when the job started running.

`bjobs -l` with a job ID displays all the information about a job, including:

- Submission parameters
- Execution environment
- Resource usage

For example:

1. Run `bjobs -l`.

bjobs -l 7678

```
Job Id <7678>, User <user1>, Project <default>, Status <PEND>, Queue <priority>, Command <verilog>
Mon Oct 28 13:08:11: Submitted from host <hostD>,CWD <$HOME>, Requested Resources <type==any &&
swp>35>;
PENDING REASONS:
Queue's resource requirements not satisfied:3 hosts;
Unable to reach slave lsbatch server: 1 host;
Not enough job slots: 1 host;

SCHEDULING PARAMETERS:
 r15s   r1m  r15m   ut   pg    io   ls  it  tmp  swp  mem
loadSched  -      0.7  1.0    -    4.0   -    -   -   -    -    -
loadStop   -      1.5  2.5    -    8.0   -    -   -   -    -    -
```

View job group information

You can view information about jobs in job groups or view jobs by job group.

- To see all job groups, run `bjgroup`.
- To see jobs by job group, run `bjobs -g /group_name`.

Monitor SLA progress

You can display the properties of service classes configured in `lsb.serviceclasses` and the dynamic state information for each service class.

1. Run `bsla`
2. Run `bacct -sla` to display historical performance of a service class.

View job output

You must be logged on as the job owner.

The output from a job is normally not available until the job is finished. However, LSF provides the `bpeek` command for you to look at the output the job has produced so far.

By default, `bpeek` shows the output from the most recently submitted job. You can also select the job by queue or execution host, or specify the job ID or job name on the command line.

To save time, you can use this command to check if your job is behaving as you expected and kill the job if it is running away or producing unusable results.

1. Run `bpeek` *job_id*.

   For example:

```
bpeek 1234
<< output from stdout >>
Starting phase 1
Phase 1 done
Calculating new parameters
...
```

View chronological history of jobs

By default, the `bhist` command displays information from the job event history file, `lsb.events`, on a per job basis.

- Use the `-t` option of `bhist` to display the events chronologically instead of grouping all events for each job.
- Use the `-T` option to select only those events within a given time range.

For example, the following displays all events which occurred between 14:00 and 14:30 on a given day:

```
bhist -t -T 14:00,14:30
Wed Oct 22 14:01:25: Job <1574> done successfully;
Wed Oct 22 14:03:09: Job <1575> submitted from host to Queue , CWD , User , Project ,
Command , Requested Resources ;
Wed Oct 22 14:03:18: Job <1575> dispatched to ;
Wed Oct 22 14:03:18: Job <1575> starting (Pid 210);
Wed Oct 22 14:03:18: Job <1575> running with execution home , Execution CWD , Execution
Pid <210>;
Wed Oct 22 14:05:06: Job <1577> submitted from host  to Queue, CWD , User , Project ,
Command , Requested Resources ;
Wed Oct 22 14:05:11: Job <1577> dispatched to ;
Wed Oct 22 14:05:11: Job <1577> starting (Pid 429);
Wed Oct 22 14:05:12: Job <1577> running with execution home, Execution CWD , Execution
Pid <429>;
Wed Oct 22 14:08:26: Job <1578> submitted from host to Queue, CWD , User , Project ,
Command;
Wed Oct 22 14:10:55: Job <1577> done successfully;
Wed Oct 22 14:16:55: Job <1578> exited;
Wed Oct 22 14:17:04: Job <1575> done successfully;
```

View history of jobs not listed in active event log

LSF periodically backs up and prunes the job history log. By default, `bhist` only displays job history from the current event log file. You can display the history for jobs that completed some time ago and are no longer listed in the active event log.

The *-n num_logfiles* option tells the `bhist` command to search through the specified number of log files instead of only searching the current log file.

Log files are searched in reverse time order. For example, the command **bhist -n 3** searches the current event log file and then the two most recent backup files.

1. Run `bhist -n` *num_logfiles*.

`bhist -n 1`	Searches the current event log file `lsb.events`
`bhist -n 2`	Searches `lsb.events` and `lsb.events.1`
`bhist -n 3`	Searches `lsb.events`, `lsb.events.1`, `lsb.events.2`
`bhist -n 0`	Searches all event log files in LSB_SHAREDIR

View job history

You can check on the status of your job since it was submitted. The `bhist` command displays a summary of the pending, suspended, and running time of jobs for the user who invoked the command.

1. Run `bhist`.
   a) Run `bhist -l` to display the time information and a complete history of scheduling events for each job.
   b) Use `bhist -u all` to display a summary for all users in the cluster.

Update interval

The job-level resource usage information is updated at a maximum frequency of every SBD_SLEEP_TIME seconds.

The update is done only if the value for the CPU time, resident memory usage, or virtual memory usage has changed by more than 10 percent from the previous update or if a new process or process group has been created.

Job-level information

Job-level information includes:

- Total CPU time consumed by all processes of a job
- Total resident memory usage in KB of all currently running processes of a job
- Total virtual memory usage in KB of all currently running processes of a job
- Currently active process group ID of a job
- Currently active processes of a job

About displaying resource allocation limits

The command blimits displays:

- Configured limit policy name
- Users (`-u` option)
- Queues (`-q` option)
- Hosts (`-m` option)
- Project names (`-P` option)

Resources that have no configured limits or no limit usage are indicated by a dash (`-`). Limits are displayed in a USED/LIMIT format. For example, if a limit of 10 slots is configured and 3 slots are in use, then `blimits` displays the limit for SLOTS as 3/10.

If limits MEM, SWP, or TMP are configured as percentages, both the limit and the amount used are displayed in MB. For example, `lshosts` displays maximum memory (maxmem) of 249 MB, and MEM is limited to 10% of available memory. If 10 MB out of are used, `blimits` displays the limit for MEM as 10/25 (10 MB USED from a 25 MB LIMIT).

Configured limits and resource usage for builtin resources (slots, mem, tmp, and swp load indices) are displayed as INTERNAL RESOURCE LIMITS separately from custom external resources, which are shown as EXTERNAL RESOURCE LIMITS.

Limits are displayed for both the vertical tabular format and the horizontal format for Limit sections. If a vertical format Limit section has no name, `blimits` displays NONAME*nnn* under the NAME column for these limits, where the unnamed limits are numbered in the order the vertical-format Limit sections appear in the `lsb.resources` file.

If a resource consumer is configured as all, the limit usage for that consumer is indicated by a dash (`-`).

PER_HOST slot limits are not displayed. The `bhosts` commands displays these as MXJ limits.

In MultiCluster, `blimits` returns the information about all limits in the local cluster.

View information about resource allocation limits

Your job may be pending because some configured resource allocation limit has been reached. You can display the dynamic counters of resource allocation limits configured in the Limit sections in `lsb.resources`.

1. Run `blimits` to display the current resource usage, including any limits that may be blocking your job.

CHAPTER 4

Platform Management Console

Platform Console

The Platform Management Console (Platform Console or Console) allows users and administrators to monitor hosts and to submit and monitor jobs.

Open the Platform Console for the first time

Launch the interactive Platform Console to monitor hosts and jobs and run jobs.

Make sure the Console is installed.

The Platform Console allows you to monitor hosts and jobs, as well as submit jobs. When logging in for the first time, the LSF administrator must log on as root.

All administrators and non-administrator users logging on to the Platform Console must provide an OS user name and password.

1. Open the Console in a supported Web browser.

   `http://master_host_name:8080/platform`
2. Log on as root.
3. If the log on failed or the URL could not be found, restart the Console by running `pmcadmin` from the command line.

   In most cases, the daemons and services required to run the Platform Console start automatically; however, if the Console goes down, you may need to restart Console services and daemons.

Enable cookies

The Platform Console requires that Web browsers accept first-party and third-party cookies. In some cases, your browser default settings may block these cookies. In this case, you need to manually change this setting.

Note:

By default, IE7 blocks cookies, while IE6, Firefox 1.5, and Firefox 2.0 allow cookies.

For IE7:

1. In your browser, go to Tools > Internet Options > Privacy > Advanced.
2. Select Override automatic cookie handling.
3. Select Accept for both first-party and third-party cookies.
4. Click OK to save your settings.

Understanding the dashboard

The dashboard is only visible to cluster and consumer administrators. If you are not logged in as a cluster or consumer administrator, you cannot see the dashboard.

Overview

The dashboard is a window into your cluster. Use the dashboard to get an overall picture of the health of your cluster and to get early warning for any systemic problems that may be occurring.

- Your hosts
- CPU utilization levels: What is right?

Note:

Your dashboard is dynamic and can change depending on what components you have integrated. For information about integration, contact Platform Support.

You can customize your tables on the dashboard and elsewhere by clicking Preferences.

Your hosts

The health of your hosts is one of the major elements of your cluster that can seriously affect its efficiency. Healthy hosts mean better throughput and a more stable production environment.

What you see

Your cluster view gives you a quick look at the health of your entire cluster at a glance. Use the Hosts by State and Hosts by CPU Utilization charts in the Cluster Health area to monitor how your hosts are doing and how hard they are working.

Beneath these charts, you can see the health of your master host and your master candidates if you have specified any. For your master host health, rest your mouse over the icon for a detailed description of its real-time attributes or click on the icon for detailed information about that particular host.

Beneath these charts, you can see the health of your master host and your master candidates if you have specified any. For your master host health, rest your mouse over the icon for a detailed description of its real-time attributes.

What to watch for

For your cluster to be healthy and working efficiently, your hosts have to be running or available to run workload units. They must also be able to run workload units at an acceptable CPU utilization level. Use the Platform Management Console to monitor for

- A large number of closed or unavailable hosts
- Hosts running at very high or very low CPU utilization

Either of these indicators means that your cluster is not running as efficiently as possible and may need some attention.

CPU utilization levels: What is right?

Only you can decide what the optimal CPU utilization level is for your workload. In some cases, 90% is acceptable and in others 70% is good. Try to recognize the level at which your hosts are attaining the highest CPU utilization level without becoming frequently unavailable.

Goal 1: Predictability and output

If you want your cluster to be as predictable as possible and for workload units to finish and produce results at a possible cost of speed, make sure your hosts are only running at 70% CPU utilization or less.

Goal 2: Fast turnaround time

If what you require is speed (you do not care if some workload units fail as long as most of them finish as quickly as possible), then 90% CPU utilization is probably the right level for you.

Host properties

The console displays both static and dynamic host properties in a separate window. You can have up to four Host Properties windows open simultaneously to compare and monitor hosts.

Some of the most useful dynamic host properties are also displayed in chart form on the Charts tab.

Host properties and descriptions are listed below. Note the different host attributes associated with a property type:

- Built-in: The host property name and definition are preconfigured and built-in to EGO. Names and definitions cannot be changed.
- Reserved: The property name is built-in to EGO, and is reserved. A reserved host attribute must be configured explicitly using this name. Definitions are configurable.
- External: Names and definitions are completely configurable. They can be redefined for other purposes. Those external attributes listed below are out-of-box host names and definitions; if the cluster administrator has reconfigured them, they may not appear on the Host Properties page.

Property	Description	Attribute type
Host Name	Name of the host.	Built-in
Status	Current state of the host: OK, Unavailable, or Closed.	Built-in
Type (Host Type)	Type of host you have. For example, LINUX86.	Built-in
CPUs (Number of CPUs)	Number of ncpus you have specified for your host.	Built-in
CPU Util	Current CPU utilization of your host in %.	Built-in
Mem (Available Memory)	Estimate of the real memory currently available to user processes. This represents the approximate size of the largest process that could be started on a host without causing the host to start paging.	Built-in

Property	Description	Attribute type
Swap (Available Swap)	Currently available virtual memory (swap space) in MB. This represents the largest process that can be started on the host (with paging).	Built-in
Pg (Paging Rate)	Virtual memory paging rate in pages per second. This index is closely tied to the amount of available memory and the total size of the processes running on a host; if there is not enough memory to satisfy all processes, the paging rate is high.	Built-in
I/O (Disk I/O Rate)	I/O throughput to disks attached directly to this host, in KB per second. This rate does not include I/O to disks that are mounted from other hosts.	Built-in
Total Slots (Total Number of Slots)	Total number of slots for this host across ALL resource groups.	Built-in
Free Slots (Number of Free Slots)	Current number of available slots for this host across ALL resource groups.	Built-in
nprocs	Number of physical processors (if ncpus is defined as procs, then ncpus=nprocs) .	Built-in
ncores	Number of cores per processor (if ncpus is defined as cores, then ncpus=nprocs x ncores).	Built-in
nthreads	Number of threads per core (if ncpus is defined as threads, then ncpus=nprocs x ncores x nthreads).	Built-in
cit	Amount of time in minutes that a CPU has been idle; configurable in `elim.sa`.	External
scvg	Resource tag identifying scavenge-ready hosts; configurable external attribute.	External
uit_t	User idle time threshold, in minutes; configurable external attribute.	External
cu_t	CPU utilization threshold, as a percentage; configurable external attribute.	External
cit_t	CPU idle time threshold, in minutes; configurable external attribute.	External
scvgf	Scavenging flag (either on or off); configurable external attribute.	External
agtctrl	When host scavenging is enabled and the scavenging agent is controlling this host, value is on.	External
15s Load (15-Second Load)	Load this host carries, averaged over the last 15 seconds. The load is the average number of processes using the CPU during a given time interval.	Built-in

Property	Description	Attribute type
15m Load (15-Minute Load)	Load this host carries, averaged over the last 15 minutes. The load is the average number of processes using the CPU during a given time interval.	Built-in
1m Load (1-Minute Load)	Load this host carries, averaged over the last minute. The load is the average number of processes using the CPU during a given time interval.	Built-in
Model (Host Model)	Model of your host. For example, Intel_EM64T.	Built-in
Process Priority	OS process priority of cluster workloads (normal or lowest).	Built-in
Host Status Reason	Reason for the current host status, if applicable.	Built-in
CPU Factor	Speed of the host's CPU relative to other hosts in the cluster. If one processor is twice the speed of another, its CPU factor should be twice as large. The CPU factors are defined by the administrator. For multiprocessor hosts, the CPU factor is the speed of a single processor; the system automatically scales the host CPU load to account for additional processors.	Built-in
Max Mem	Maximum RAM available.	Built-in
Max Swap	Maximum swap space on your host.	Built-in
Temp (Available Temp)	Space available in MB on the file system that contains the temporary directory.	Built-in
Max Temp	Maximum space in `/tmp` (Linux/UNIX) or OS default temp directory (Windows).	Built-in
Disks	Number of local disks on your host.	Built-in
It (Idle Time)	Amount of time in minutes that a host has been idle. On a Linux/UNIX host, it is the amount of time since the keyboard has been touched on all logged in sessions. On a Windows host, it is the amount of time a screen saver has been active.	Built-in
Users (Login Users)	Number of current users logged in to the system.	Built-in
Resource Attr	Resource attributes assigned to this host. For example, "mg" indicates the host is a management host; "scvg" indicates the host is scavenge-ready.	Built-in
CPU Speed	Speed of each individual CPU in MHz.	Built-in
Band Width	Maximum bandwidth requirement in Mbps.	Built-in

Charts

You can see the host charts by clicking on the name or icon of a particular host. When the Host Properties window opens, click the Charts tab.

The host properties charts give you streaming data and recent historical data (up to 30 minutes) about your host performance. Use these charts to troubleshoot a host or keep an eye on a host that may be performing poorly. Keep the host properties page open to collect data up to 30 minutes.

Chart	Description
CPU Utilization	A real-time measurement of CPU utilization of your host in percentage.
Total Available Memory	A real-time indicator of the current available memory in MB on your host.
Paging Rate	A real-time indicator of the virtual memory paging rate in pages per second.
Available Swap	A real-time indicator of the current available swap in MB on your host.
15-Second Load	A real-time measurement of the load of your computer averaged over the last 15 seconds. The load is the average number of processes using the CPU during a given time interval.
Disk I/O Rate	A real-time measurement of the I/O throughput to disks attached directly to this host, in KB per second. This rate does not include I/O to disks that are mounted from other hosts.

Host dashboard

Using the Platform Management Console's host dashboard view, you can see the state of all your hosts.

Host icon	What it means
	A green base with green bars means that your host is OK and that it is running workload units. The more CPU capacity a host is using, the higher the green bars. A green outline means that the host has been activated at the lowest priority (for host scavenging).
	A grey icon means that this host is closed. It may be closed for several reasons.
	A red icon means that the host is unavailable.

The master host icon and name are displayed at the bottom of the Console page.

How do I see details?

For any host, hold your mouse over the icon for details about the current state of the host. For even more detail, click on the host icon. The Host Properties page opens with all the details for that host.

Monitor hosts and services

Monitor host states

From the Hosts page, you can see the status of monitored hosts along with additional information.

From the Charts page, you can choose to generate Charts for any host or collection of hosts in the cluster. For example, you can view Charts that illustrate compute node loads or memory utilization on the primary install node.

1. From the organization tree, select Resources > Hosts.

   The icon view of all hosts in your cluster displays and identifies your current master host.
2. Hover over any host icon.

   Details about the host display including host name and current status.
3. Click any host icon.

   The host details page displays.
4. Click Charts and generate the graphs you want to display.

Create a system report

You can create reports that summarize trends, events, alerts, and the availability of a range of services. From here you can also view notifications and event logs.

Although many graphs, reports, and charts are automatically generated within the Platform Console, you can also use the reporting function to customize and build specific reports that illustrate host and service trends and availability, along with alert statistics and summaries.

1. From the organizational tree, open the System Monitoring folder and then click Reporting.
2. Select the report type you want to build from the options at the top of the page.

   Choose from Trends, Availability, Alert Histogram, or Alert Summary.

   The corresponding Reporting page opens.
3. Follow the steps and instructions provided on the Reporting page, and then click Continue until all information is provided for report generation.

Monitor your cluster: present and past performance

Platform RTM must be installed on an LSF cluster (Update 3 and above) before integrating with the Platform Console.

Install Platform RTM to allow a graphical view of your LSF cluster. Platform RTM communicates the overall health of multiple LSF clusters, as well as data about past cluster performance.

Once integrated with the Platform Console, you can use RTM to create reports and view cluster details in the same way as you do using the RTM Console. See the *Platform RTM User Guide* for details.

Note:

General steps for integrating Platform RTM with the Platform Console are provided. Check the related Platform LSF and Platform RTM documents for details.

1. Enable the PMC plug-in for the RTM Console.

   See the *Platform LSF Command Reference* for information on running `pmcadmin addrtm` *URL* to integrate the RTM web pages.
2. From the organizational tree in the Platform Console, click RTM.
3. Provide the URL to the RTM Console.

   For example: `http://host_name/cacti`.

   Platform RTM options appear in the Platform Console organizational tree.

Manage jobs and run work

General information

The Platform Console allows you to submit and monitor jobs, and provides access to job reports. You can choose to submit jobs directly from the Platform Console, or through a separate workload management application (for example, Platform Lava or Platform LSF). After logging on to the Platform Console, you can monitor the jobs you submitted through either the Console or a workload management application.

There are a number of standard interfaces in the Platform Console through which you can submit jobs. A generic interface is provided, along with several pre-configured and customizable interfaces (such as Fluent, LS-DYNA, ABAQUS, ANSYS, Nastran, and EnginFrame).

Generic job submission interface

The Generic interface allows you to submit LSF jobs via a web browser. The Generic interface cannot be customized.

Submit jobs by running a binary or script as a job command (for example, `/bin/sleep 10`). Use this interface to specify most `bsub` options including input, output, and error files, notification and queue details, processor and resource requirements, and host and job limits.

Other job submission interfaces

By default, the Platform Console provides some templates for several popular job submission interfaces. You can add new ones, and customize existing interfaces to suit your workload submission requirements. Included interfaces are briefly described here.

Fluent

This is an application-specific interface for job submission. Fluent must be installed to use this interface. The out-of-box interface can be customized.

Submit CAT, DAT, and journal files that reside locally or on a server, and provide parameters and job options that are specific to the Fluent application, such as time stamp and iteration number.

To integrate with the actual Fluent application, you must change the `FLUENT_CMD` to point to the application For example:

```
FLUENT_CMD="/pcc/app/fluent/Fluent.Inc/bin/fluent"
```

Find the script file here:

```
LSF_TOP/gui/lsf/7.0/batchgui/plugin/lsf/exec/fluent.cmd
```

LS-DYNA

This is an application-specific interface for job submission. LS-DYNA must be installed to use this interface. The out-of-box interface can be customized.

Submit d3dump files that reside locally or on a server, point to the input deck, and provide parameters and job options that are specific to the LS-DYNA application.

To integrate with the actual LS-DYNA application, you must change the `DYNA_CMD` to point to the application For example:

```
DYNA_CMD="$DYNA_TOP/binaries/ls_dyna/linux/$VERSION"
```

Find the script file here:

```
LSF_TOP/gui/lsf/7.0/batchgui/plugin/lsf/exec/lsdyna.cmd
```

ABAQUS

This is an application-specific interface for job submission. ABAQUS must be installed to use this interface. The out-of-box interface can be customized.

Submit INP, restart, and user files that reside locally or on a server, and provide parameters and job options that are specific to the ABAQUS application.

To integrate with the actual ABAQUS application, you must change the `ABAQUS_CMD` to point to the application For example:

```
ABAQUS_CMD="/pcc/app/abaqus/6.5.3/x86-64/Commands/abaqus"
```

Find the script file here:

```
LSF_TOP/gui/lsf/7.0/batchgui/plugin/lsf/exec/abaqus.cmd
```

ANSYS

This is an application-specific interface for job submission. ANSYS must be installed to use this interface. The out-of-box interface can be customized.

Submit computing (.db), load case, and TXT files that reside locally or on a server, and provide parameters that are specific to the ANSYS application, such as the parallel type and MPP solver.

To integrate with the actual ANSYS application, you must change the `ANSYS_CMD` to point to the application For example:

```
ANSYS_CMD="/pcc/app/ansys_inc/v110/ansys/bin/ansys110"
```

Find the script file here:

```
LSF_TOP/gui/lsf/7.0/batchgui/plugin/lsf/exec/ansys.cmd
```

Nastran

This is an application-specific interface for job submission. Nastran must be installed to use this interface. The out-of-box interface can be customized.

Point to an input deck and provide parameters that are specific to the Nastran application.

To integrate with the actual Nastran application, you must change the `NASTRAN_CMD` to point to the application For example:

```
NASTRAN_CMD="/pcc/app/msc/nastran/2005/bin/msc2005 nastran"
```

Find the script file here:

```
LSF_TOP/gui/lsf/7.0/batchgui/plugin/lsf/exec/nastran.cmd
```

EnginFrame

Through the Platform Console, you can submit and monitor workload using an external interface. EnginFrame must be installed to take advantage of this feature.

Open an EnginFrame job submission interface

To access an EnginFrame job submission interface from the Platform Console, you must provide its location.

1. From the organizational tree, open Manage Workload > Submission.

   A list of standard interfaces displays in the tree.
2. Click EnginFrame.
3. Enter the location of the installed EnginFrame application.

   For example, `http://host_name:port/enginframe`.
4. Click Enable.

Submit a job using an interface

This procedure uses the Generic interface as an example of how to submit a job through the Platform Console.

1. From the organizational tree, open Manage Workload > Submission.

   A list of standard interfaces displays in the tree from which you can submit jobs using desired submission options.
2. Click an interface (for example, Generic).
3. From the job submission page, provide the required job details (for example, job name, optional commands, input file, host and job limits, etc.)
4. Click Submit Job.

Manage a submitted job

Once a job is submitted, you can manage the job from the Platform Console. Management options include killing (stopping) jobs, suspending jobs, or resuming jobs.

1. From the organizational tree, open the Manage Workload folder and click Batch Jobs.

   The Jobs page opens and lists all jobs active on the cluster.
2. Select one or more jobs from the list, and then click Kill, Suspend, or Resume from the top of the page, as required.

View job details

You can view details about current and recent jobs in the cluster (by default, completed jobs are purged from the system after one hour).

1. From the organizational tree, open Manage Workload > Batch Jobs.

   The Jobs page opens and lists all jobs active on the cluster.
2. From the list, click a job ID.

   A page with additional details about the job displays.

View or customize LSF workload reports

The Reports function requires Platform LSF 7 installed in the cluster. Note that the primary install node must be an LSF node.

You can view standard and custom LSF workload data using the Platform Console, or customize LSF reports.

1. From the organizational tree, open Manage Workload > Reports.

2. For the database configuration parameters specified in LSF, see your Platform LSF installation documents for details.
3. To view LSF reports:
   - Click Standard Reports to view standard reports.
   - Click Custom Reports to view existing custom reports.
4. To customize LSF reports:
   1. Click Custom Reports.

      The Reports (List) page opens, listing any existing customized reports.
   2. Click the Global Actions drop-down list and select Create Custom Report.
   3. Enter the required fields, and then click Create.

      The new customized report type displays on the Reports (List) page.

To understand the LSF reporting interface, see *Administering Platform LSF* for details.

Index

L

M

N

O

P

Q

R

S

T

U