KEK System Blue Gene Solution

IBM Japan
ITRO&HPC Services
Fumiyasu Ishibashi

Summary

- **Blue Gene @ KEK**
- **KEK Blue Gene System Overview**
- **Job Scheduling**
  - LoadLeveler + Wrapper script
  - Job script file
  - Queue (job-classes)
  - Accounting
  - First-In First-Out (FIFO)
  - Inter-group equalization
  - Notification
  - Performance report
- **Monitoring Fatal errors**
- **Utilization**

Blue Gene @ KEK

KEK

Kou **E**nerugi kasokuki **K**enkyuu kikou (Japanese)

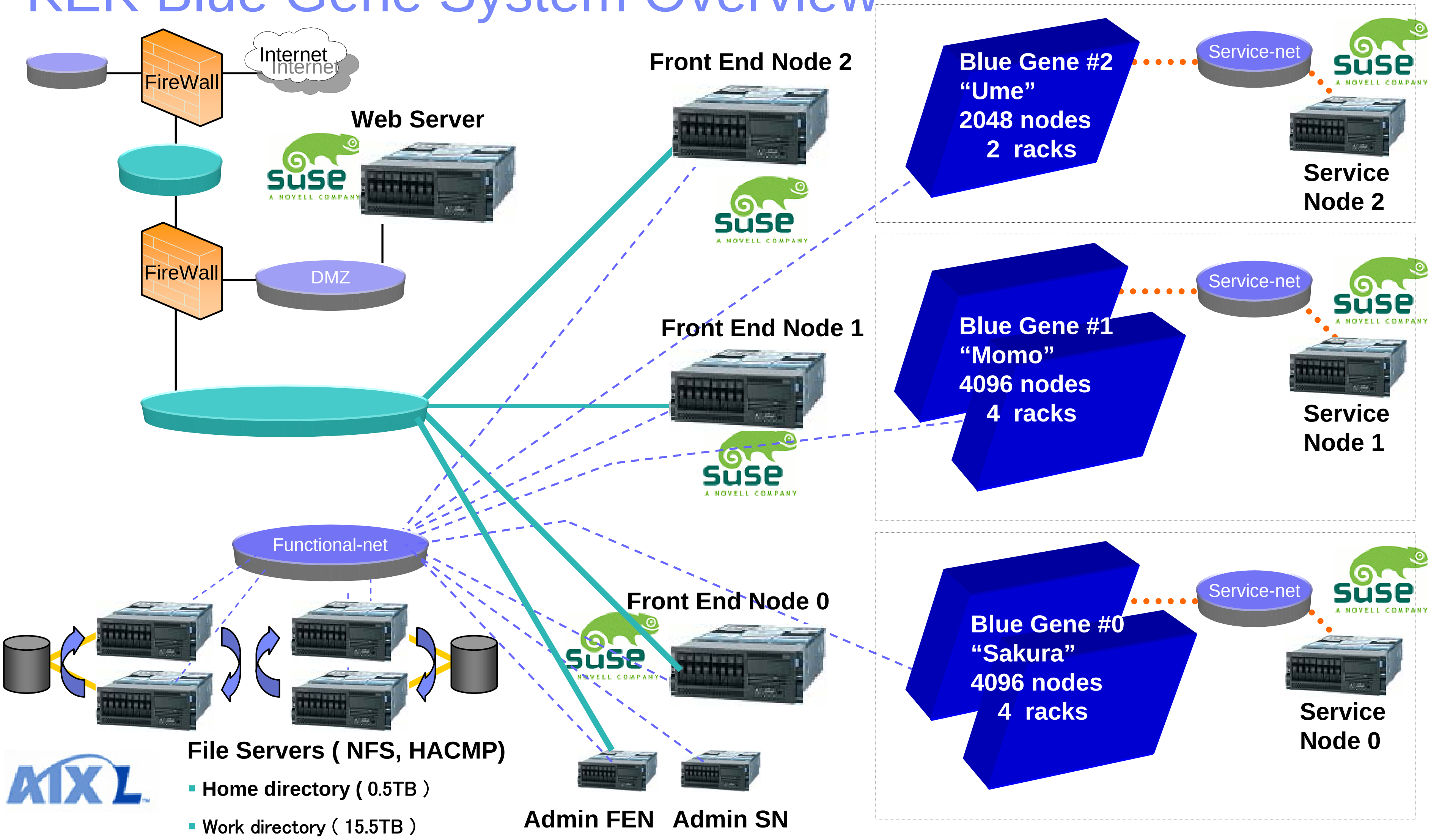
High Energy Accelerator Research Organization (English)

http://www.kek.jp/intra-e/index.html

Blue Gene @ KEK

- Installation: 05' Dec~06' Feb @KEK
- Service started: 06' March 1 10:00~
- 10 Racks (10,240 c-nodes)
- Peak Performance: 57.3TFLOPS

KEK Blue Gene System Overview

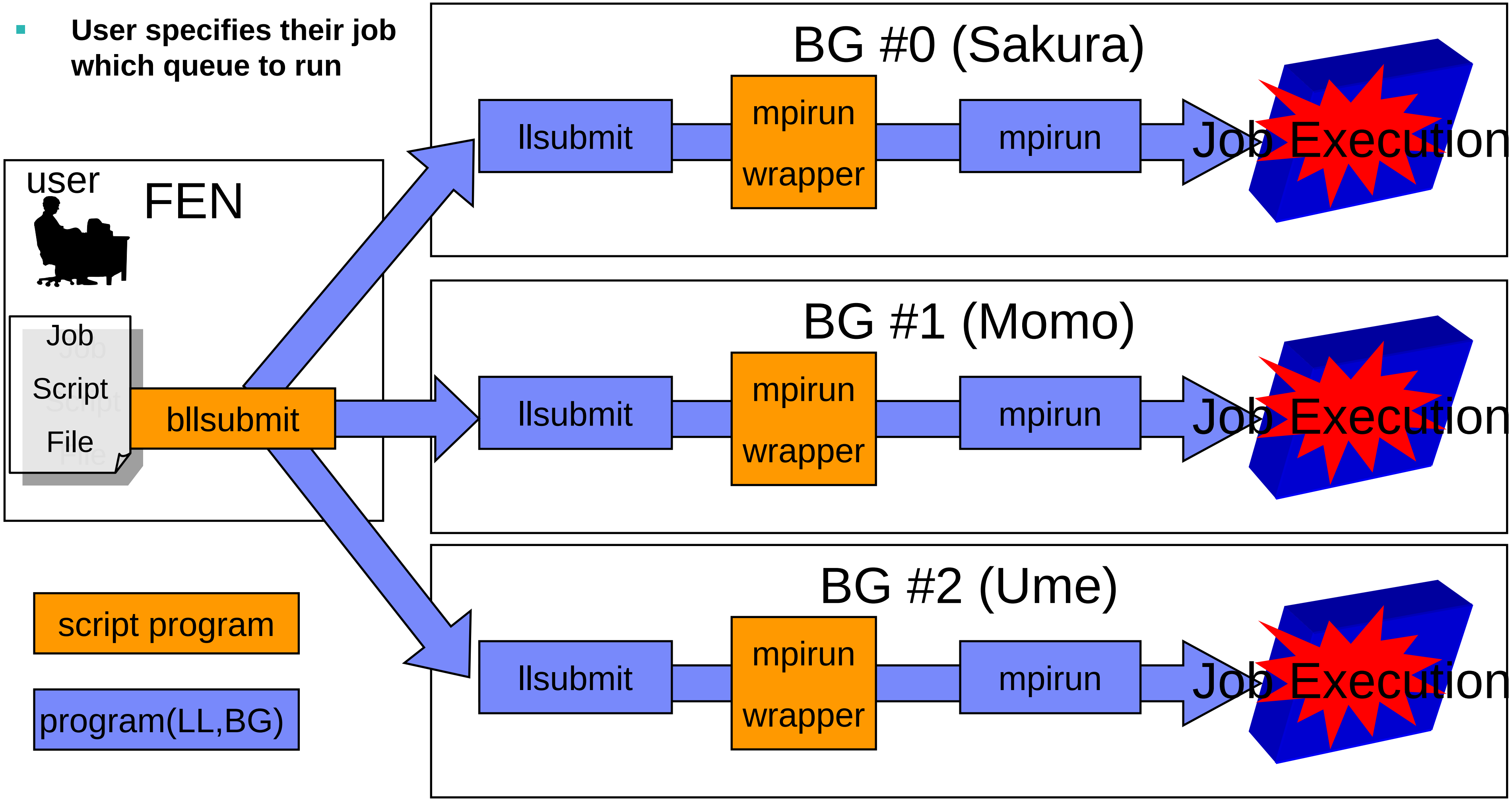


File Servers (NFS, HACMP)

- Home directory (0.5TB)
- Work directory (15.5TB)

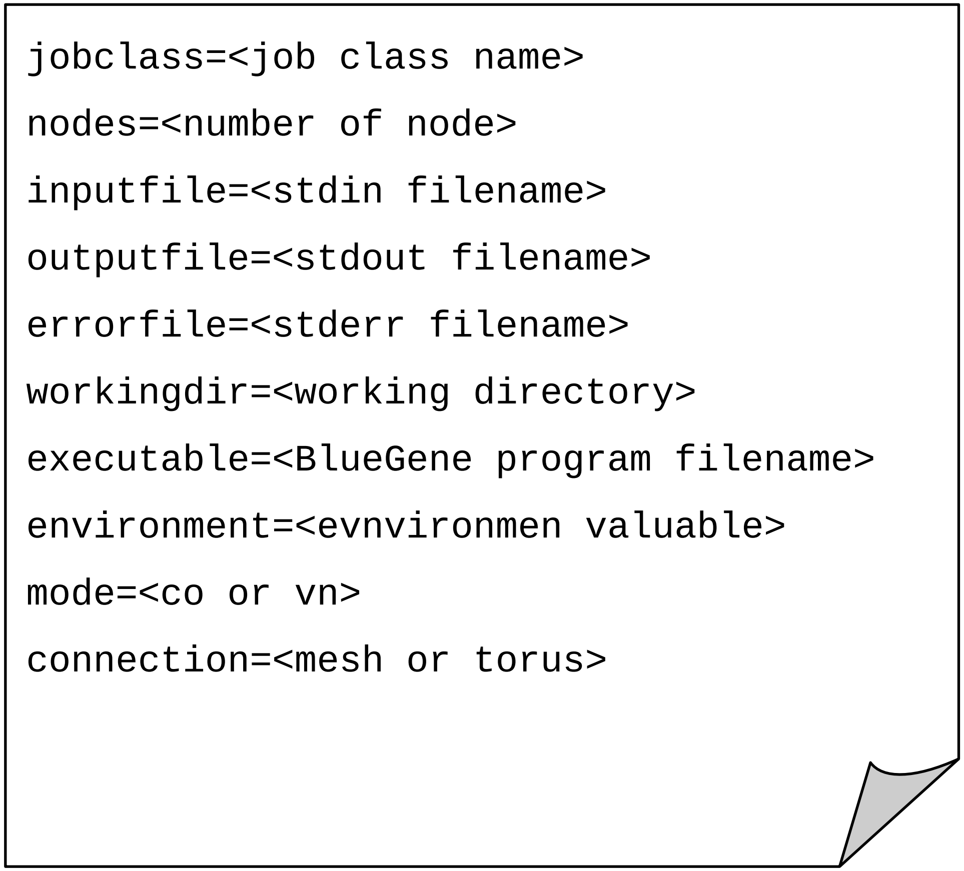
Job Scheduling: LoadLeveler + Wrapper script

- **User specifies their job which queue to run**



Job Scheduling: Job script file

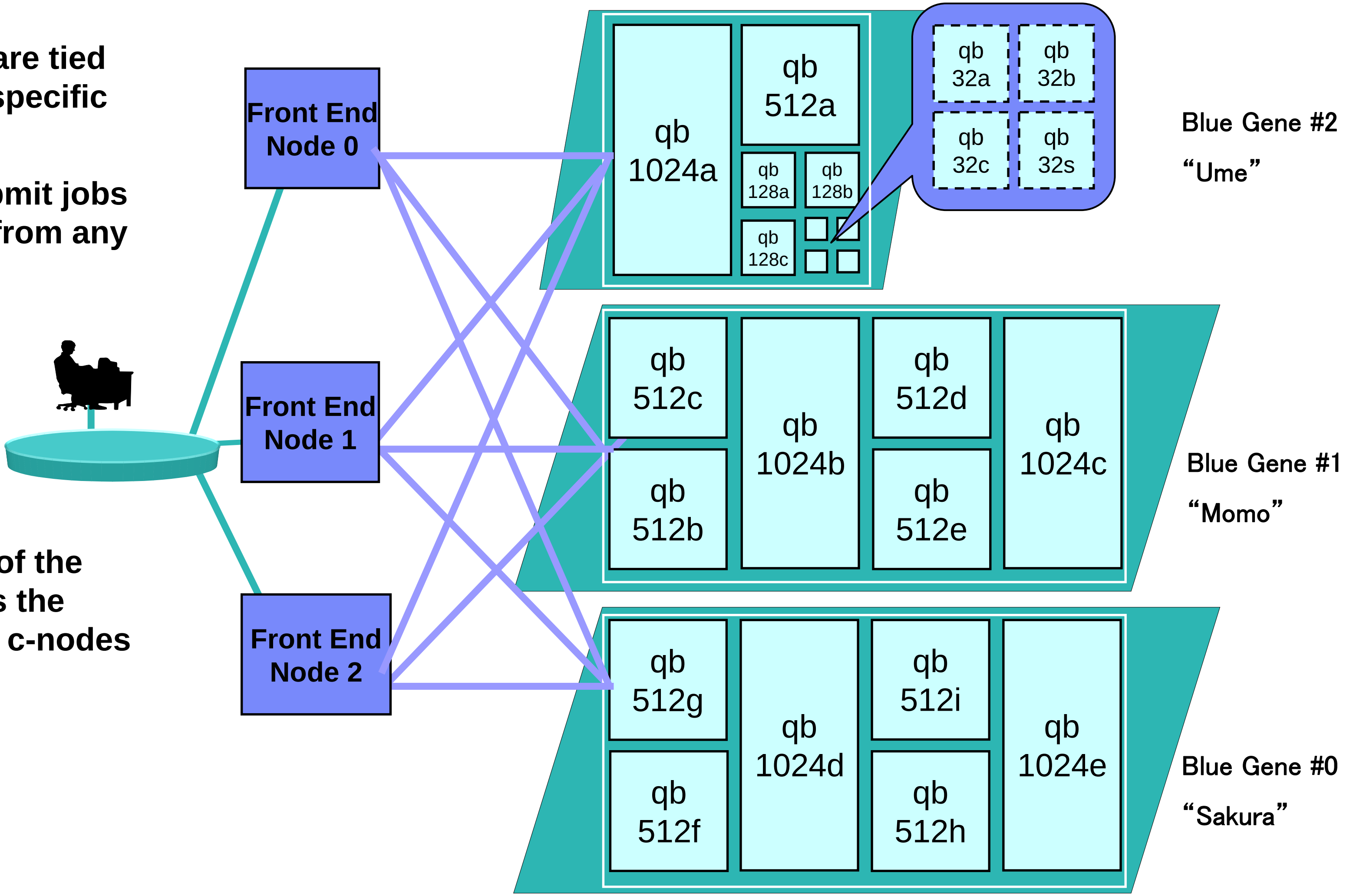
```
$ bllsubmit [job script file]
```



```
jobclass=<job class name>
nodes=<number of node>
inputfile=<stdin filename>
outputfile=<stdout filename>
errorfile=<stderr filename>
workingdir=<working directory>
executable=<BlueGene program filename>
environment=<evnvironmen valuable>
mode=<co or vn>
connection=<mesh or torus>
```

Job Scheduling: Queue (job-class)

- **Each queues are tied together to a specific block**
- **Users can submit jobs to any queue from any FEN**
- **The numbers of the queue name is the number of the c-nodes**



Job Scheduling: Accounting

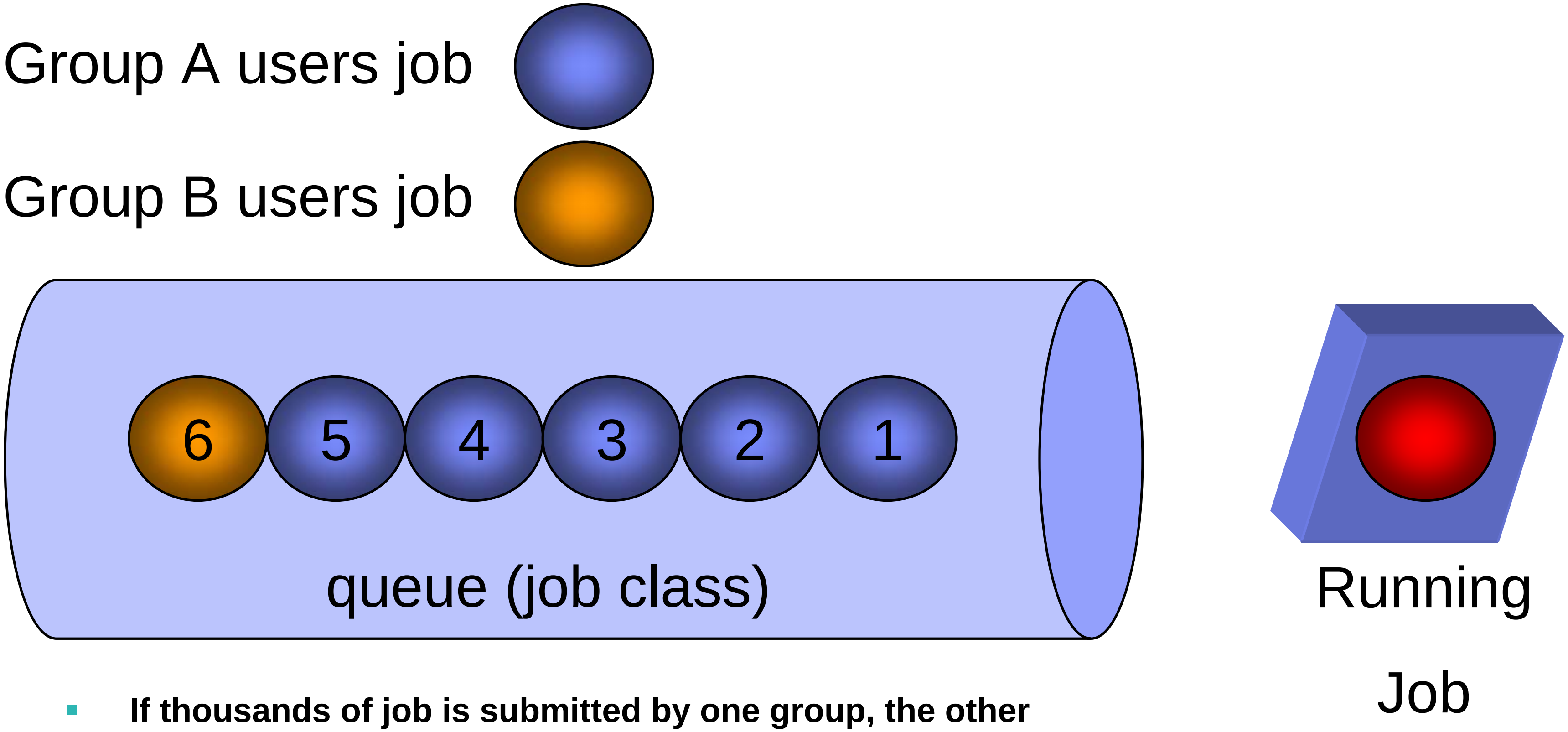
- **Billed per Unix group**
- **Billed per job-group**
- **User can check his/her accounting information by a script (bgtlst)**
- **Users of the group who used up the limit time will not be able to submit new jobs**

```
ibm-fumi@b0fe0ad:~> bgtlst
    USER        JG         LIMIT         USED         LEFT
--------        --     ---------    ---------    ---------
ibm-fumi        SS             -    000003:33    000006:26
ibm-fumi         S             -    000001:20    000000:00
ibm-fumi         M             -    000000:00    000008:41
ibm-fumi         L             -    000006:46    000003:14
--------        --     ---------    ---------    ---------
   GROUP        JG         LIMIT         USED         LEFT
--------        --     ---------    ---------    ---------
  scbadm        SS     000010:00    000003:34    000006:26
  scbadm         S     000010:00    000001:20    000008:40
  scbadm         M     000010:00    000001:19    000008:41
  scbadm         L     000010:00    000006:46    000003:14
--------        --     ---------    ---------    ---------
```

Job-group Definition

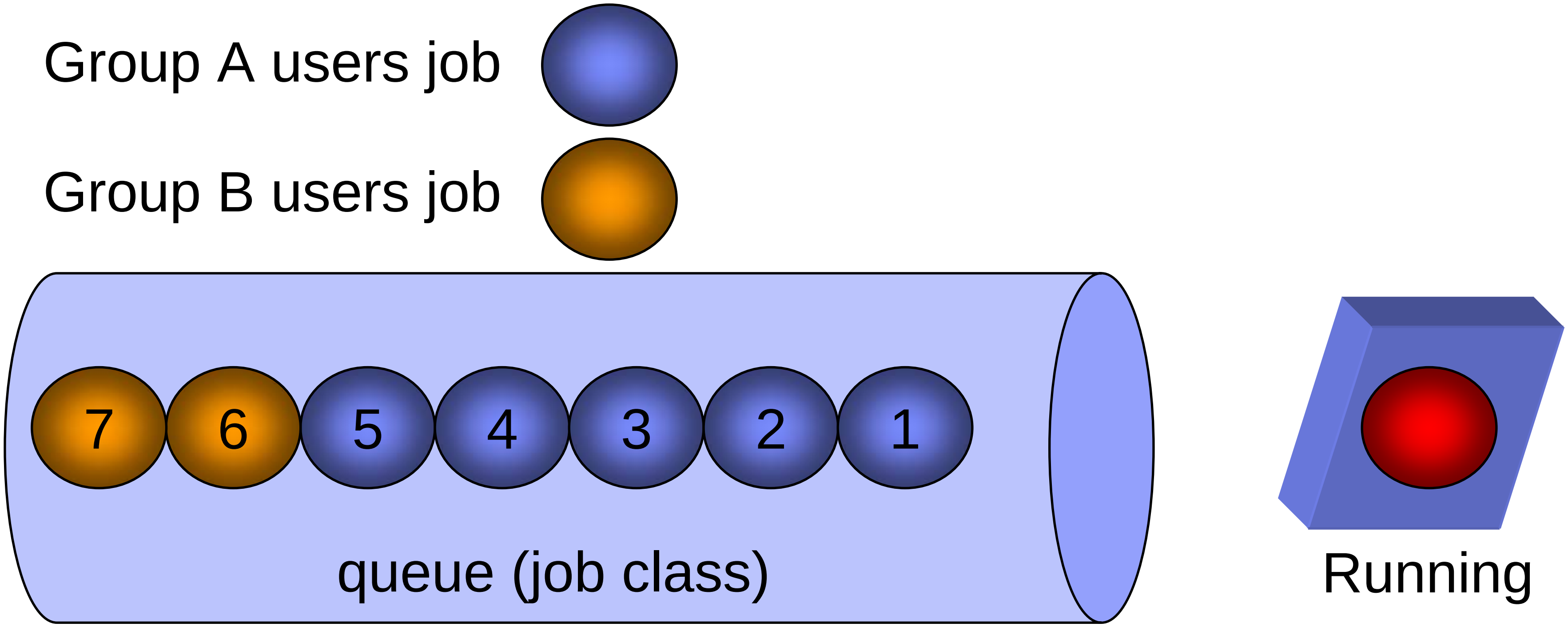
```
SS: 32   nodes (qb32x)
S : 128  nodes (qb128x)
M : 512  nodes (qb512x)
L : 1024 nodes (qb1024x)
```

Job Scheduling: First-In First-Out (FIFO)



- **If thousands of job is submitted by one group, the other group will not be able to run the job**

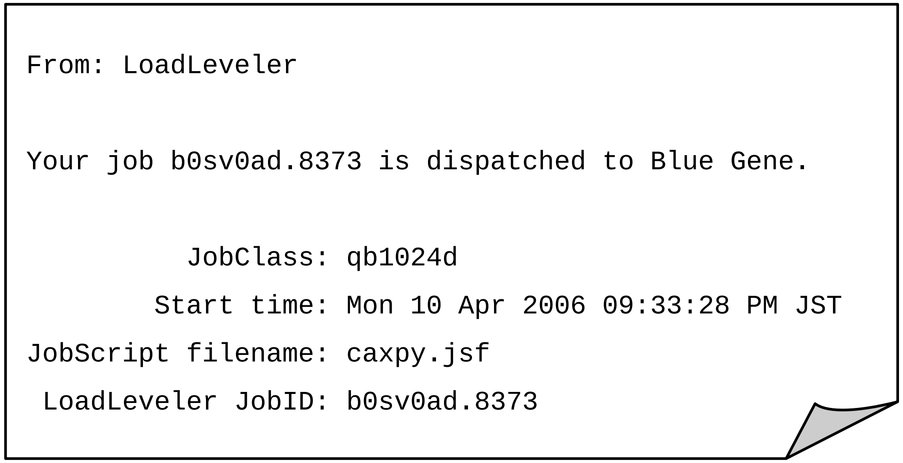
Job Scheduling: Inter-group equalization



- **The structure is made by job submitting wrapper script and LoadLeveler configuration.**

Job Scheduling: Notification

- **User can specify if he/she wants to be notified**
  - **when the job starts**
  - **when the job ends**
  - **where to e-mail**
- **Mailed by a mpirun-wrapper script**



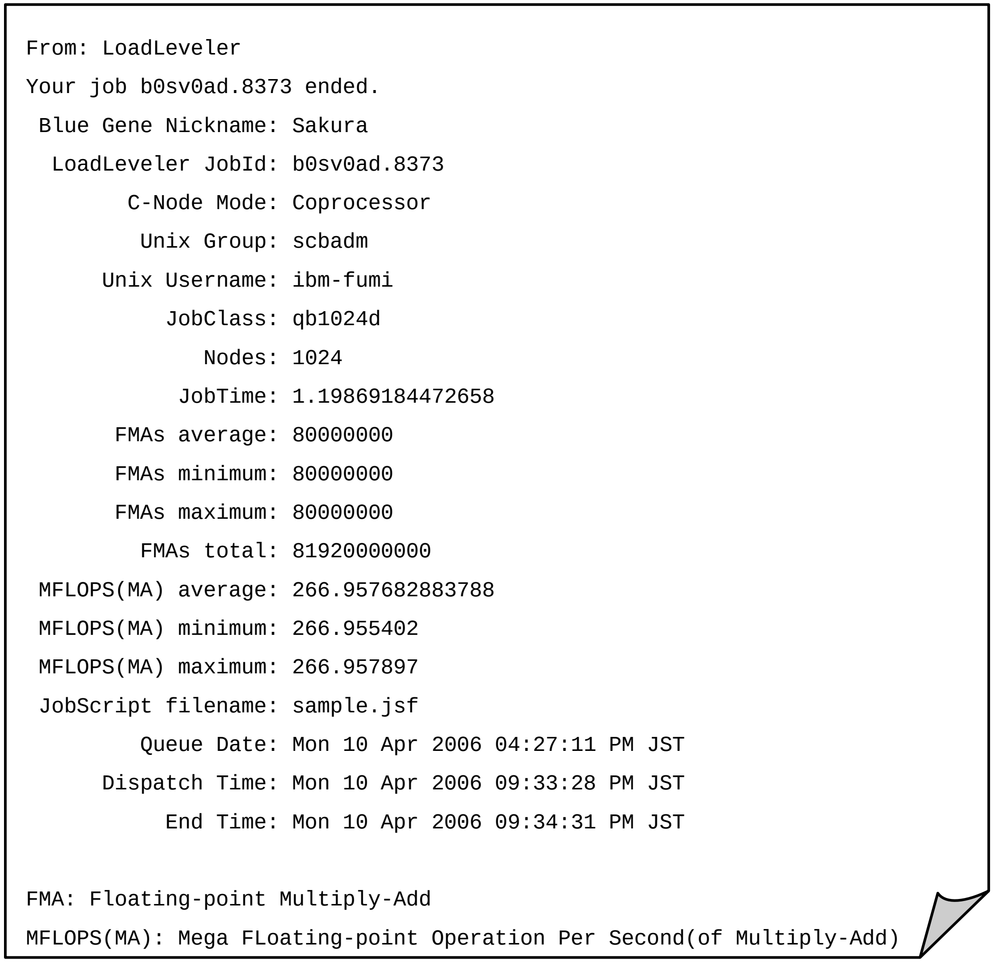
```
From: LoadLeveler

Your job b0sv0ad.8373 is dispatched to Blue Gene.

          JobClass: qb1024d
        Start time: Mon 10 Apr 2006 09:33:28 PM JST
JobScript filename: caxpy.jsf
 LoadLeveler JobID: b0sv0ad.8373
```

Job Scheduling: Performance report

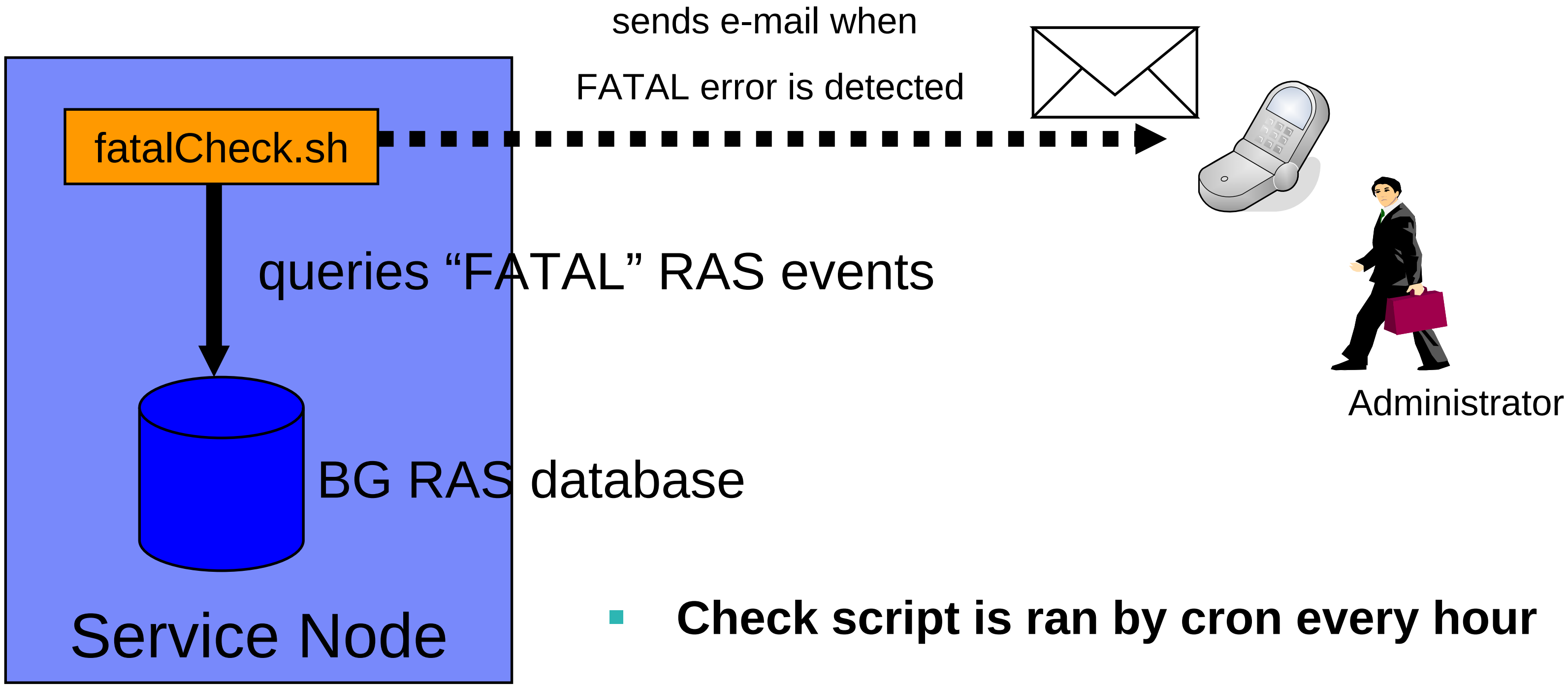
- **Reported in the job completation mail (if the mail address not specified, a file will be made under the users home directory)**
- **Formated and send by a mpirun-wrapper script**



```
From: LoadLeveler
Your job b0sv0ad.8373 ended.
 Blue Gene Nickname: Sakura
  LoadLeveler JobId: b0sv0ad.8373
        C-Node Mode: Coprocessor
         Unix Group: scbadm
      Unix Username: ibm-fumi
           JobClass: qb1024d
              Nodes: 1024
            JobTime: 1.19869184472658
       FMAs average: 80000000
       FMAs minimum: 80000000
       FMAs maximum: 80000000
         FMAs total: 81920000000
 MFLOPS(MA) average: 266.957682883788
 MFLOPS(MA) minimum: 266.955402
 MFLOPS(MA) maximum: 266.957897
 JobScript filename: sample.jsf
         Queue Date: Mon 10 Apr 2006 04:27:11 PM JST
      Dispatch Time: Mon 10 Apr 2006 09:33:28 PM JST
           End Time: Mon 10 Apr 2006 09:34:31 PM JST

FMA: Floating-point Multiply-Add
MFLOPS(MA): Mega FLoating-point Operation Per Second(of Multiply-Add)
```

Monitoring Fatal errors

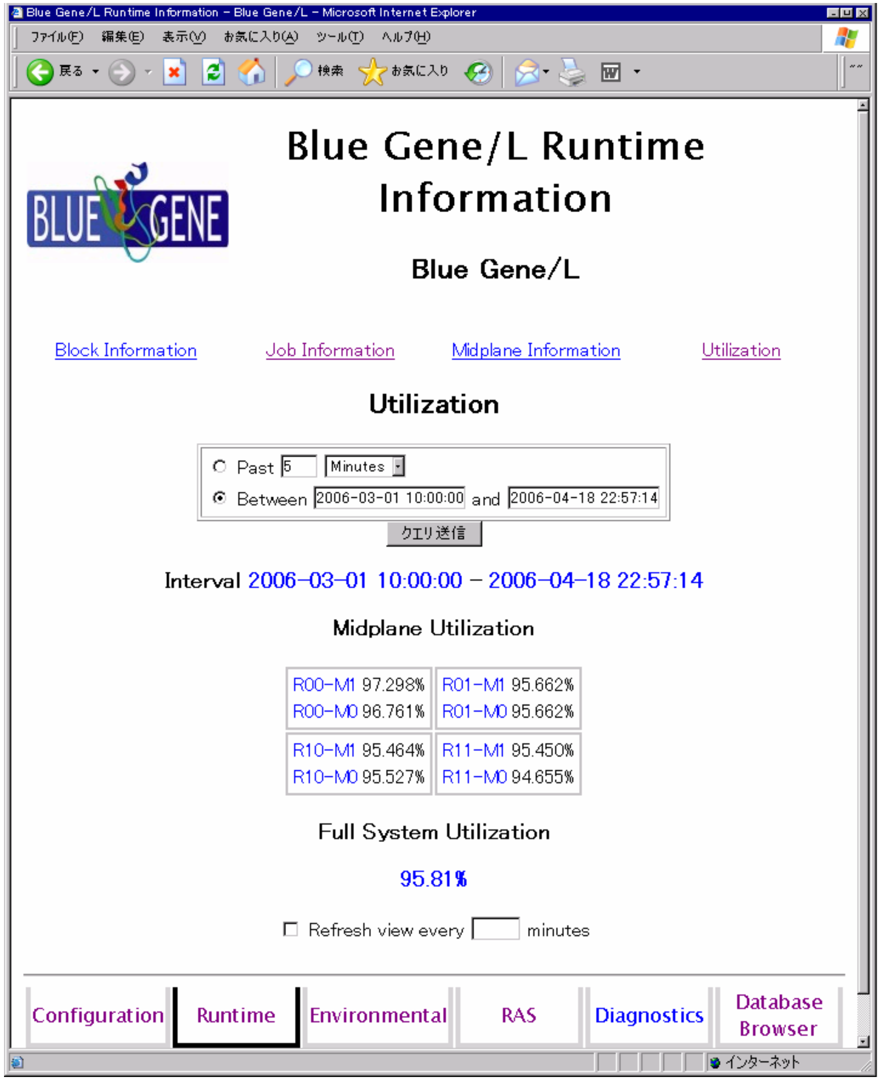


- **Check script is ran by cron every hour**
- **Planning to use Trigger function of DB2**

Utilization

- **Usually 100~500 jobs queued on each Blue Gene (Ume, Momo, Sakura)**
- **Utilization (2006 3/1 – 4/18) from BG runtime database**
  - 85%(full system) *
  - 95%(blocks used for 512-1024nodes job)*

  * job times that was terminated by defect, user reason are included
- **Utilization should increase with more users and less defects**


Blue Gene/L Runtime Information - Blue Gene/L - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
BLUE GENE
インターネット

Blue Gene/L Runtime Information

Blue Gene/L

Block Information | Job Information | Midplane Information | Utilization

Utilization

○ Past 5 Minutes
● Between 2006-03-01 10:00:00 and 2006-04-18 22:57:14

クエリ送信

Interval 2006-03-01 10:00:00 – 2006-04-18 22:57:14

Midplane Utilization

R00-M1 97.298%	R01-M1 95.662%
R00-M0 96.761%	R01-M0 95.662%
R10-M1 95.464%	R11-M1 95.450%
R10-M0 95.527%	R11-M0 94.655%

Full System Utilization

95.81%

☐ Refresh view every ___ minutes

Configuration | Runtime | Environmental | RAS | Diagnostics | Database Browser

Thankyou!