

IBM at ISC19

IBM Spectrum LSF & Scale User Group

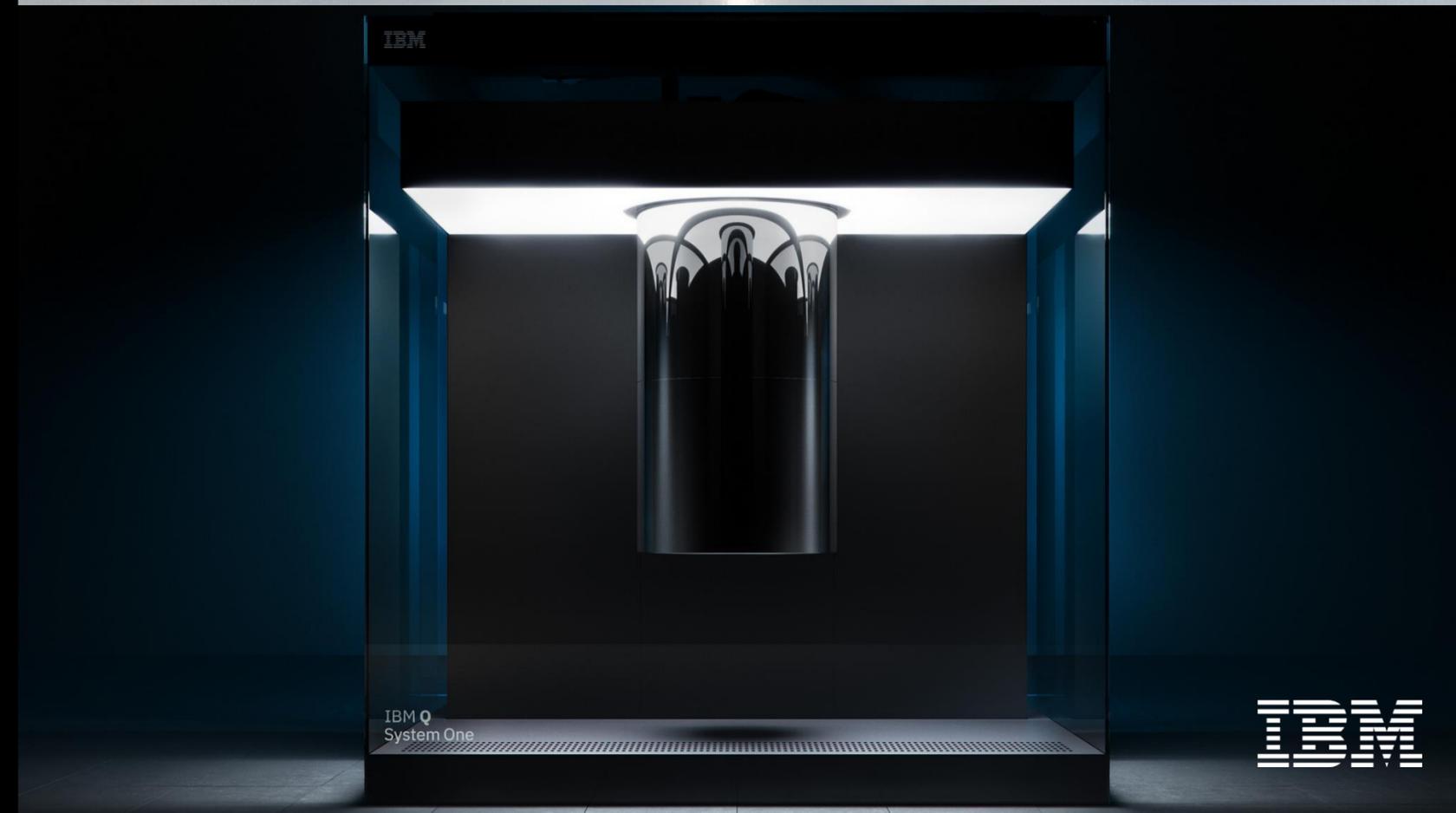
IBM Spectrum LSF Update

Bill.McMillan@uk.ibm.com

Global Offering Leader

IBM Spectrum LSF

IBM Cognitive Systems



Notices and disclaimers

© 2019 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information.

This document is distributed “as is” without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity. IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply.”

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer’s responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products.

IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

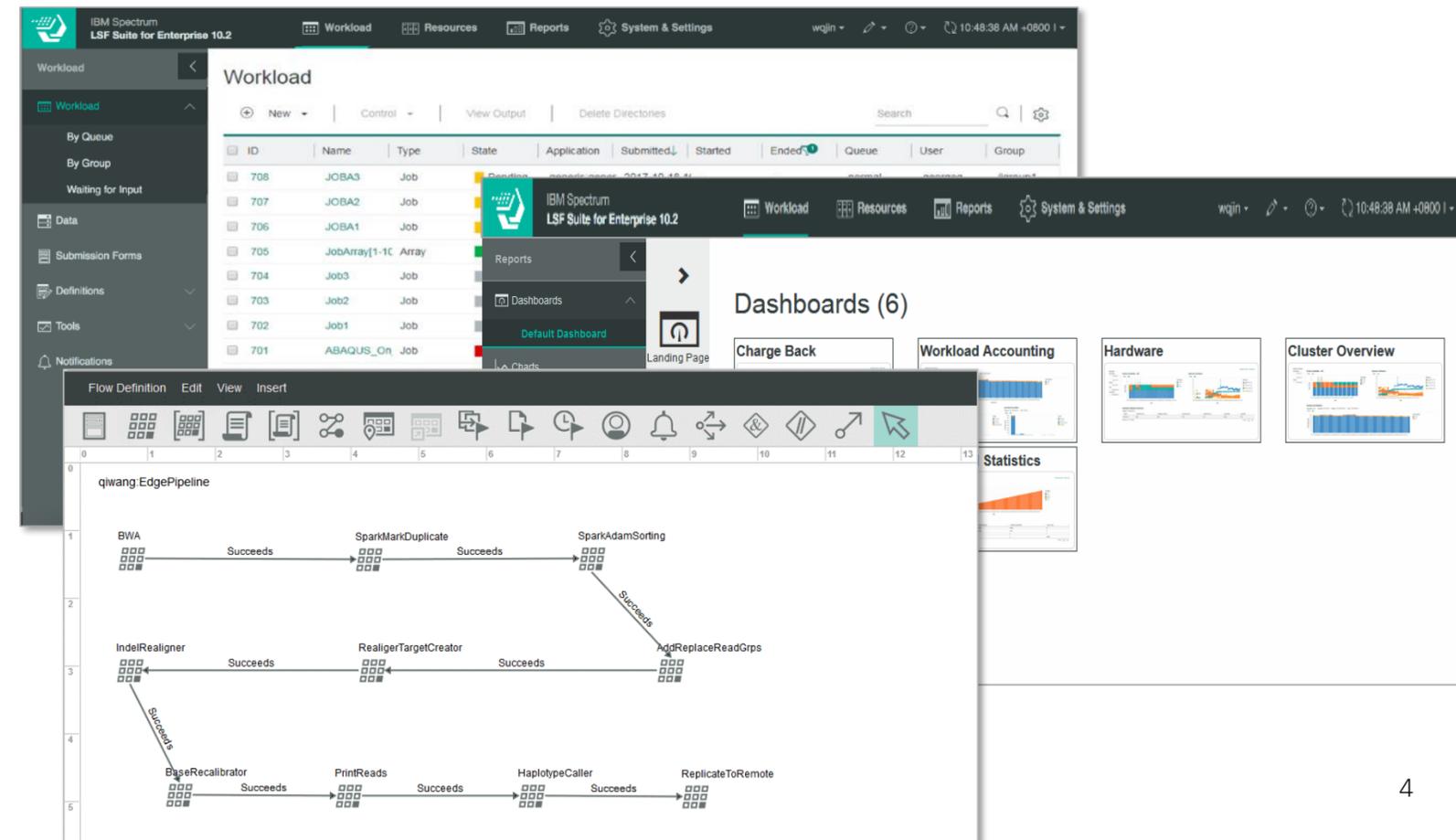
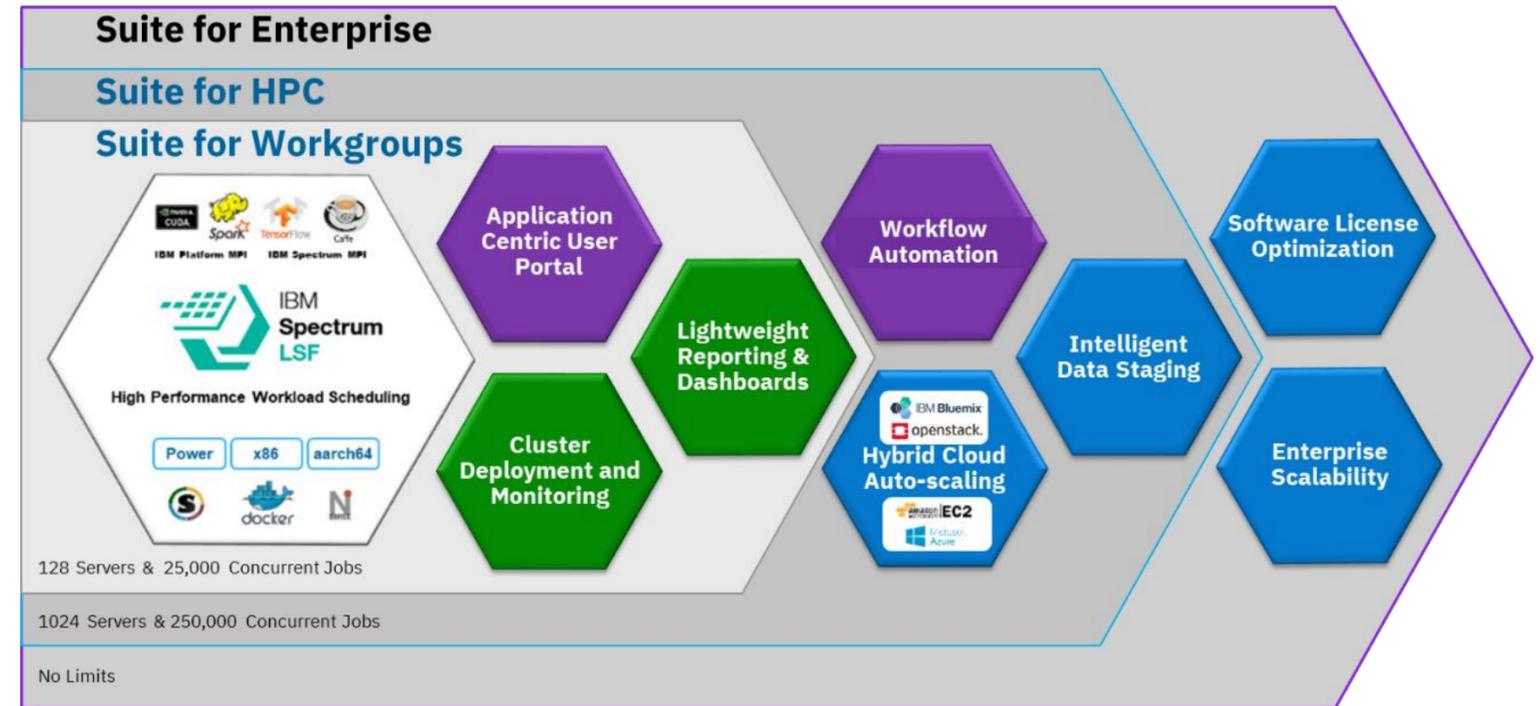
IBM Spectrum LSF Suite

Everything you need for Workgroup, HPC or Enterprise scale.

- **Enhanced Utilization of assets through effective scheduling and sharing policies**
- **Enhancing User Productivity through ease of use, accessibility and simplification**
- **Operational Efficiency through insight of how the HPC environment is being used**

The LSF Suite for HPC is available at no charge via the IBM Academic Initiative

Hourly Pricing now available.



Simplify the User Experience

Workload

⊕ New ▾ | Control ▾ | View Output | Delete Directories | Search 🔍 | ⚙️

<input type="checkbox"/>	ID	Name	Type	State	Application	Submitted↓	Started	Ended	Queue	User	Group
<input type="checkbox"/>	708	JOBA3	Job	 Pending	generic:gener	2017-10-18 16:00	-	-	normal	georgeg	/jgroup1
<input type="checkbox"/>	707	JOBA2	Job	 Pending	generic:gener	2017-10-18 16:00	-	-	normal	georgeg	/jgroup1
<input type="checkbox"/>	706	JOBA1	Job	 Pending	generic:gener	2017-10-18 16:00	-	-	normal	georgeg	/jgroup1
<input type="checkbox"/>	705	JobArray[1-10]	Array		generic:gener	2017-10-18 16:00	2017-10-18 16:00	-	normal	georgeg	-
<input type="checkbox"/>	704	Job3	Job	 Done	generic:gener	2017-10-18 16:00	2017-10-18 16:00	2017-10-18 16:00	normal	georgeg	-
<input type="checkbox"/>	703	Job2	Job	 Done	generic:gener	2017-10-18 16:00	2017-10-18 16:00	2017-10-18 16:00	normal	georgeg	-
<input type="checkbox"/>	702	Job1	Job	 Done	generic:gener	2017-10-18 16:00	2017-10-18 16:00	2017-10-18 16:00	normal	georgeg	-
<input type="checkbox"/>	701	ABAQUS_On	Job	 Exited	ABAQUS:AB/	2017-10-18 16:00	2017-10-18 16:00	2017-10-18 16:00	normal	georgeg	-
<input type="checkbox"/>	700	tt	Job	 Exited	-	2017-10-17 16:00	2017-10-17 16:00	2017-10-17 16:00	normal	georgeg	-
<input type="checkbox"/>	699	ABAQUS_150	Job	 Done	ABAQUS:AB/	2017-10-16 16:00	2017-10-16 16:00	2017-10-16 16:00	normal	georgeg	-

Job slot limit reached: 4 hosts

Page 2 of 3 | 10 ▾ | Viewing 11 - 20 of 27

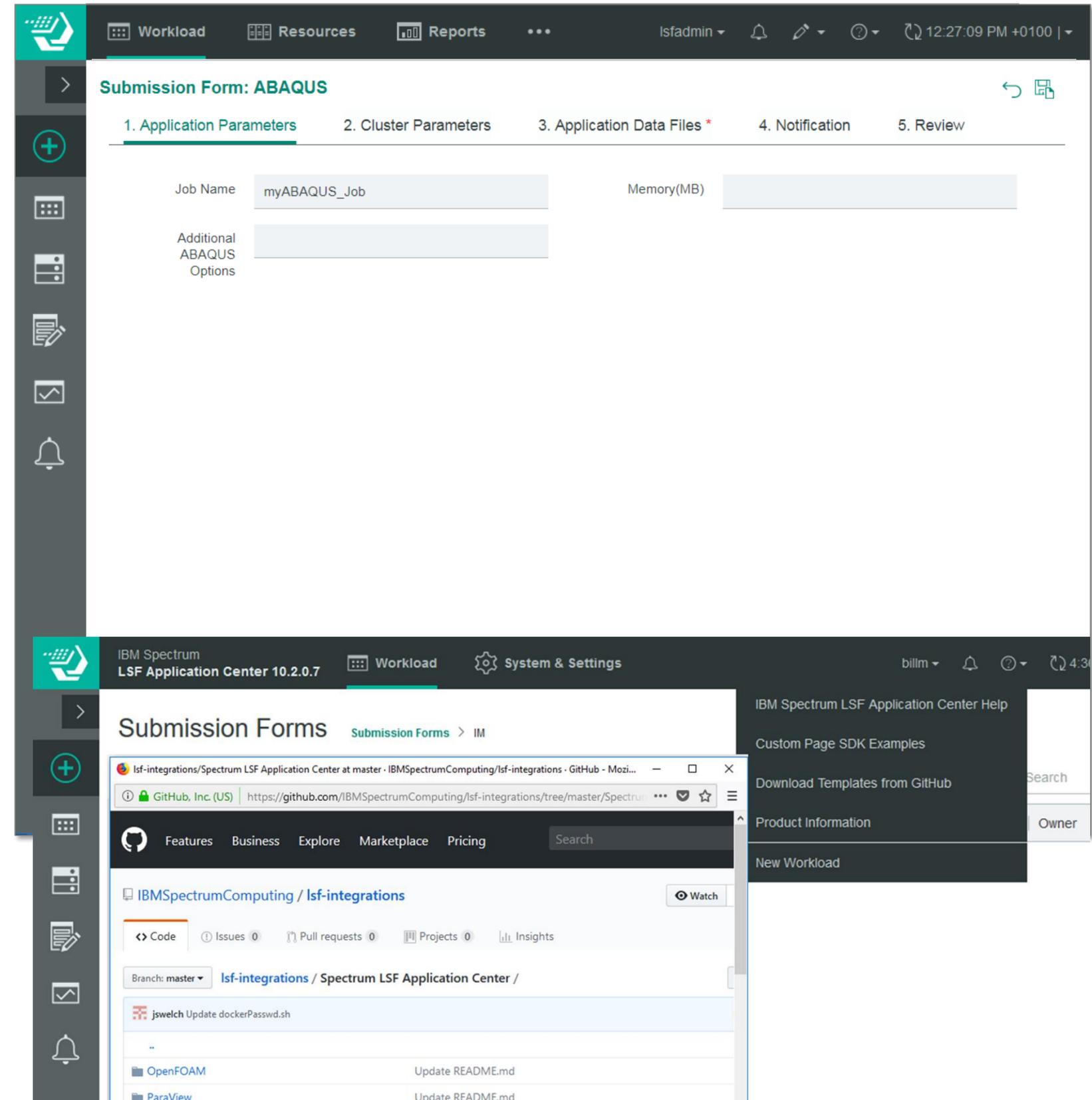
Guided Submission

The administrator or power user can define templates for specific applications.

- Display as a form or wizard (new in SPK8)

Templates:

- Simplify the use of the application exposing specific fields with (optional) predefined options/values.
- Reduces user training requirements, errors and support user
- Makes HPC more accessible – point and click submission
- Templates can handle dependencies and define how input/output files for an application should be viewed/graphed.
- Download additional/updated templates from Github



Example – Containerised Tensorflow + Tensorboard

The screenshot displays the LSF Suite for HPC 10.2.0.6 Workload management interface. The browser address bar shows the URL: <https://gauss01:8443/platform/#/framework/workload/workload/workloadlist>. The top navigation bar includes tabs for Workload, Resources, System & Settings, and Reports. The user is logged in as student3, and the time is 10:26:29 AM -0700.

The main content area is titled "Workload" and features a search bar and several action buttons: New, Control, View Output, and Delete Directories. The search filters are set to "User = student3" and "Ended = Past Hour". Below the filters is a table with the following columns: ID, Type, Name, State, Submitted, Started, Ended, and User. The table currently displays "No data to display".

ID	Type	Name	State	Submitted	Started	Ended	User
No data to display							

The bottom of the page shows a JavaScript error message: "javascript:;".

Simplifying Viewing of Results

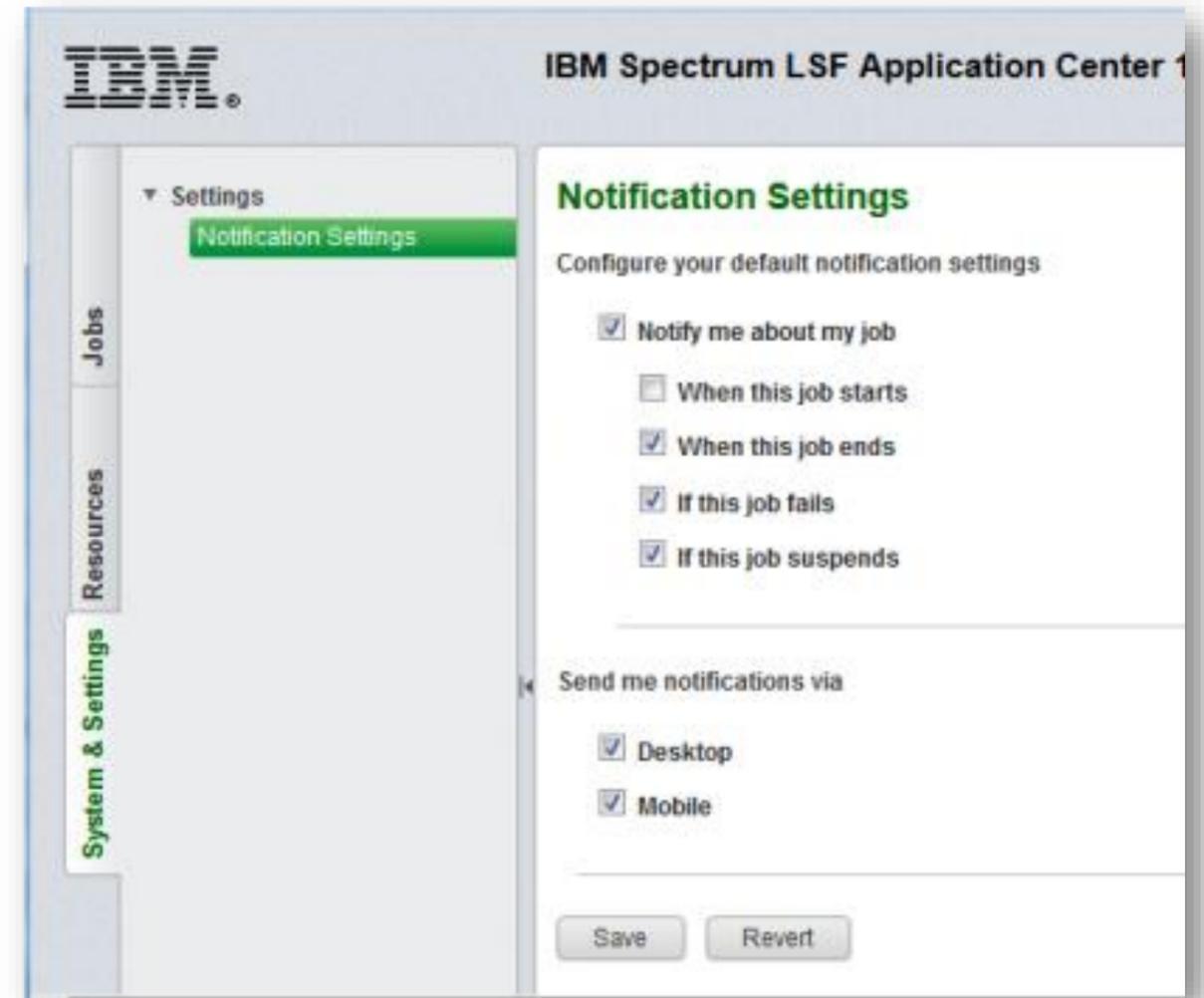
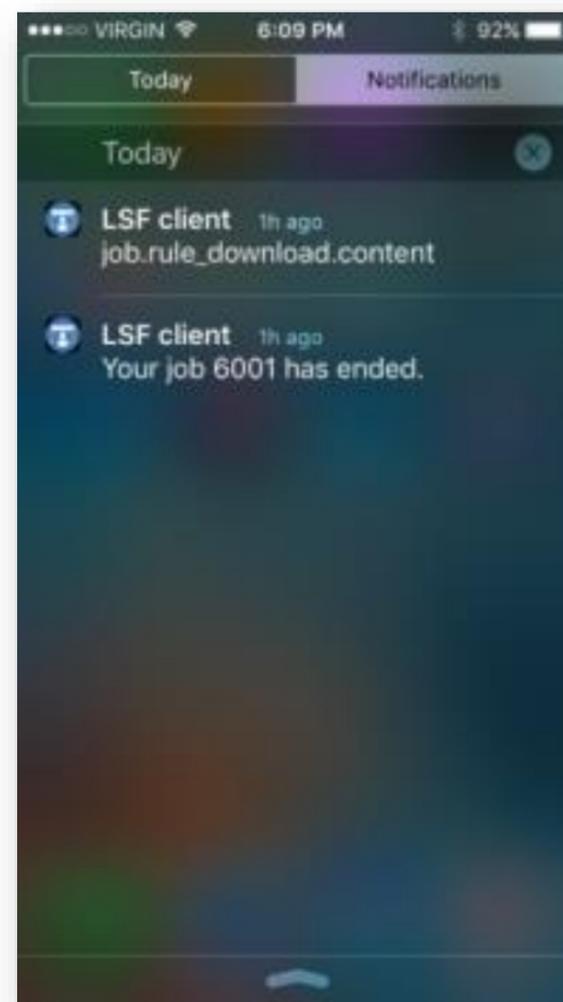
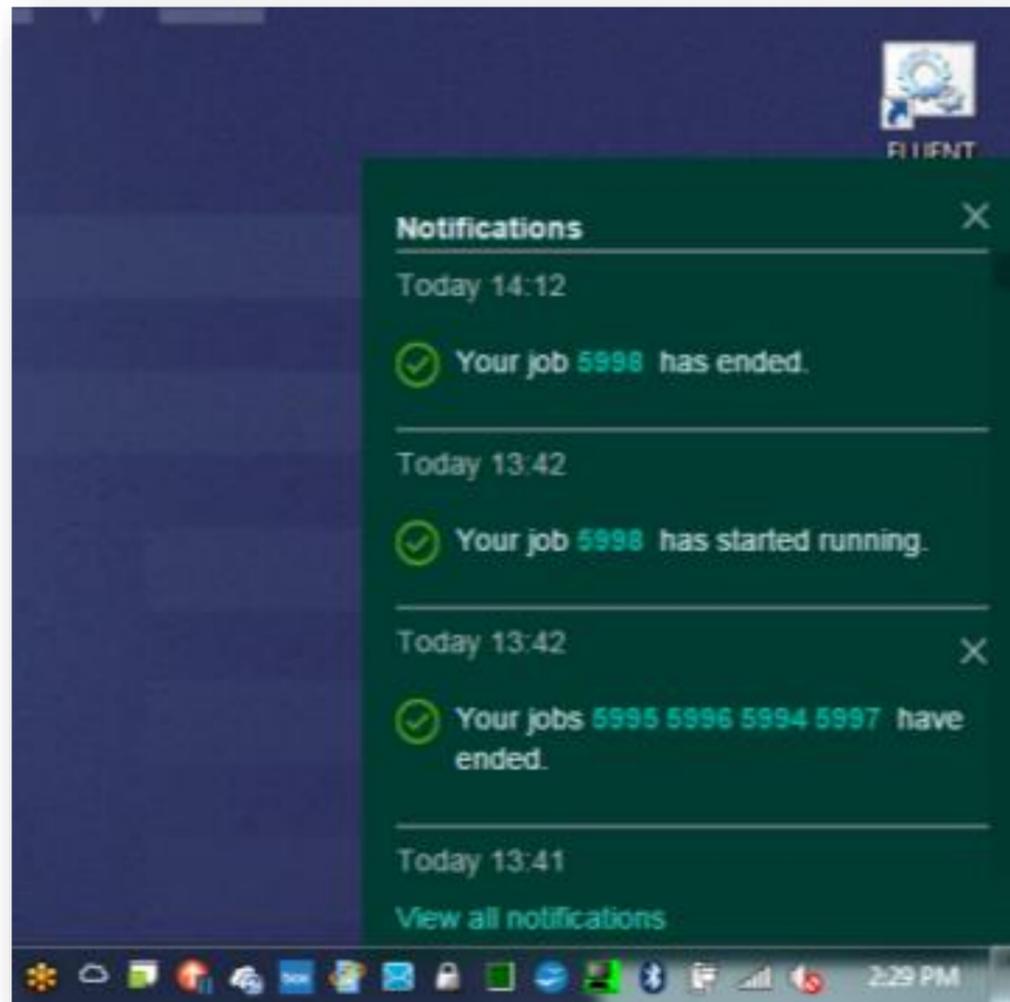
The output of most applications are one or more text files which you then have to post process to turn into something useful.

The image displays several windows from a data analysis application:

- File Browser:** Shows a list of files under the job name 'ABAQUS_JOB100', including 'output.georgeg.txt' and 'abaqus-test.inp'.
- Tail Window:** Displays the last 30 lines of a text file. The data is as follows:

Line	Time	Value 1	Value 2	Value 3	Value 4	Value 5
1	258	0.000326	0.002708	0.000339	0.002569	0.000583
2	259	0.000316	0.002599	0.000330	0.002352	0.000567
3	260	0.000308	0.002530	0.000320	0.003175	0.000551
4	261	0.000298	0.002469	0.000310	0.002419	0.000536
5	262	0.000290	0.002355	0.000300	0.002260	0.000520
6	263	0.000282	0.002270	0.000292	0.002360	0.000504
7	264	0.000275	0.002202	0.000281	0.002207	0.000487
8	265	0.000268	0.002077	0.000269	0.002142	0.000471
9	266	0.000260	0.001973	0.000260	0.001982	0.000454
10	267	0.000255	0.001929	0.000251	0.003109	0.000438
11	268	0.000247	0.001826	0.000242	0.002132	0.000422
12	269	0.000239	0.001737	0.000234	0.001890	0.000406
13	270	0.000231	0.001658	0.000227	0.002095	0.000390
14	271	0.000225	0.001612	0.000215	0.002860	0.000374
15	272	0.000215	0.001581	0.000206	0.001943	0.000358
16	273	0.000206	0.001493	0.000200	0.001962	0.000342
17	274	0.000199	0.001432	0.000194	0.002266	0.000326
18	275	0.000191	0.001390	0.000187	0.001715	0.000310
19	276	0.000182	0.001320	0.000181	0.001912	0.000294
20	277	0.000176	0.001262	0.000174	0.001667	0.000278
21	278	0.000169	0.001242	0.000169	0.002503	0.000262
22	279	0.000162	0.001226	0.000163	0.001882	0.000246
23	280	0.000154	0.001176	0.000159	0.001710	0.000230
24	281	0.000147	0.001125	0.000157	0.001581	0.000214
25	282	0.000143	0.001136	0.000154	0.002242	0.000198
- Execution Period Chart:** A bar chart showing three categories: SYSTEM TIME (SEC) at approximately 14, TOTAL CPU TIME (SEC) at approximately 36, and WALLCLOCK TIME (SEC) at approximately 21.
- Data Chart (Line Graph):** Shows multiple data series over time, with values decreasing from approximately 0.0003 to 0.0001.
- 3D Surface Plot:** A 3D plot of a surface with axes labeled 'z', 'Uy', and 'Epsilon Type'. The surface is colored according to a scale from 0.2 (blue) to 1.0 (red).

Inform the user of the progress, completion or failure of their work



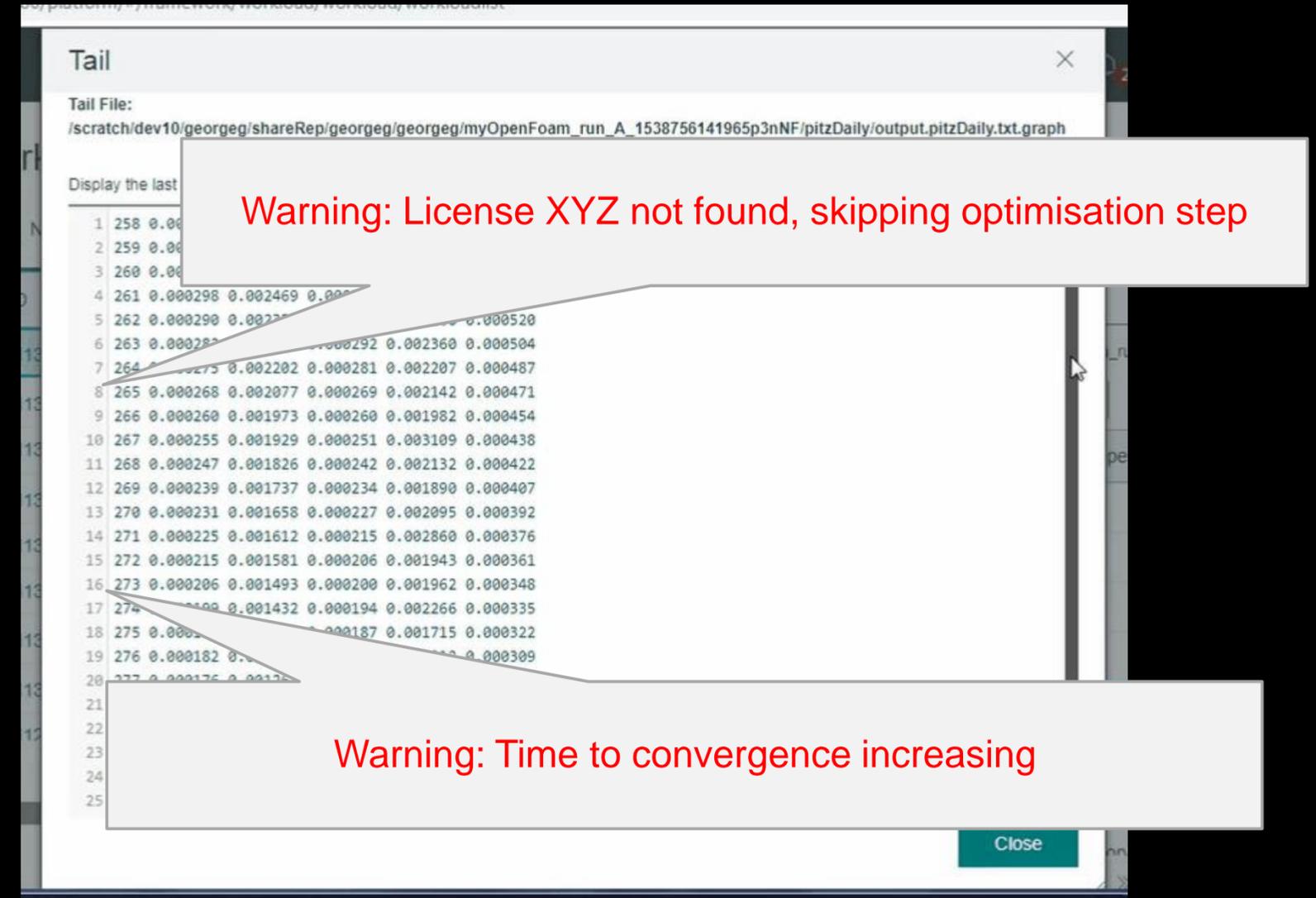
Email, desktop, browser or mobile

But what about jobs that are still running, but are not quite “right”?

Users repeatedly check the output of applications to check that the application is doing the right thing.

There may be something wrong, but it hasn't caused the application to fail – such as a solution not converging.

If the user doesn't notice this, the job may run for many hours and produce no useful output.



New Application Specific Watchdog

The screenshot shows the IBM Spectrum LSF Application Center 10.2.0.8 interface. The main 'Workload' section contains a table with the following data:

ID	Name	State	External Status	Contai
17776	Longjob	Running	High Memory Usage (500M)	-
18079	sleep 600	Exited	-	-
18080	sleep 111	Done	-	-
18081	sleep 500	Exited	Job hang, killed by system.	-
18082	sleep 200	Exited	Job hang, killed by system.	-
18083	with_Watchdog	Exited	High memory useage 106 M	-
18084	sleep 101	-	-	-

The 'Notifications' panel shows two alerts for Job <17776>: High Memory Usage (500M). An email alert is also shown:

Job warning for <17776:Longjob>
 noreply to: George Gao

From: noreply@ib22b11.localdomain
 To: George Gao/Ontario/IBM@IBMCA

High Memory Usage (500M)

The following is the workload details:

ID: [17776](#)
 Name: Longjob
 Application: generic:generic
 Started: 2019-05-09 11:41:33

The screenshot shows an iPhone notification center with the following alerts:

- LSF client 1h ago: job.rule_download.content
- LSF client 1h ago: Your job 6001 has ended.
- LSF Client Job 11319: Fatal "Optimisation failed. Aborted"
- LSF Client Job 11321: Warning "Solution time diverging"

IBM & ASTON MARTIN RED BULL RACING



Focus Areas 2019-2021

Core Scheduling

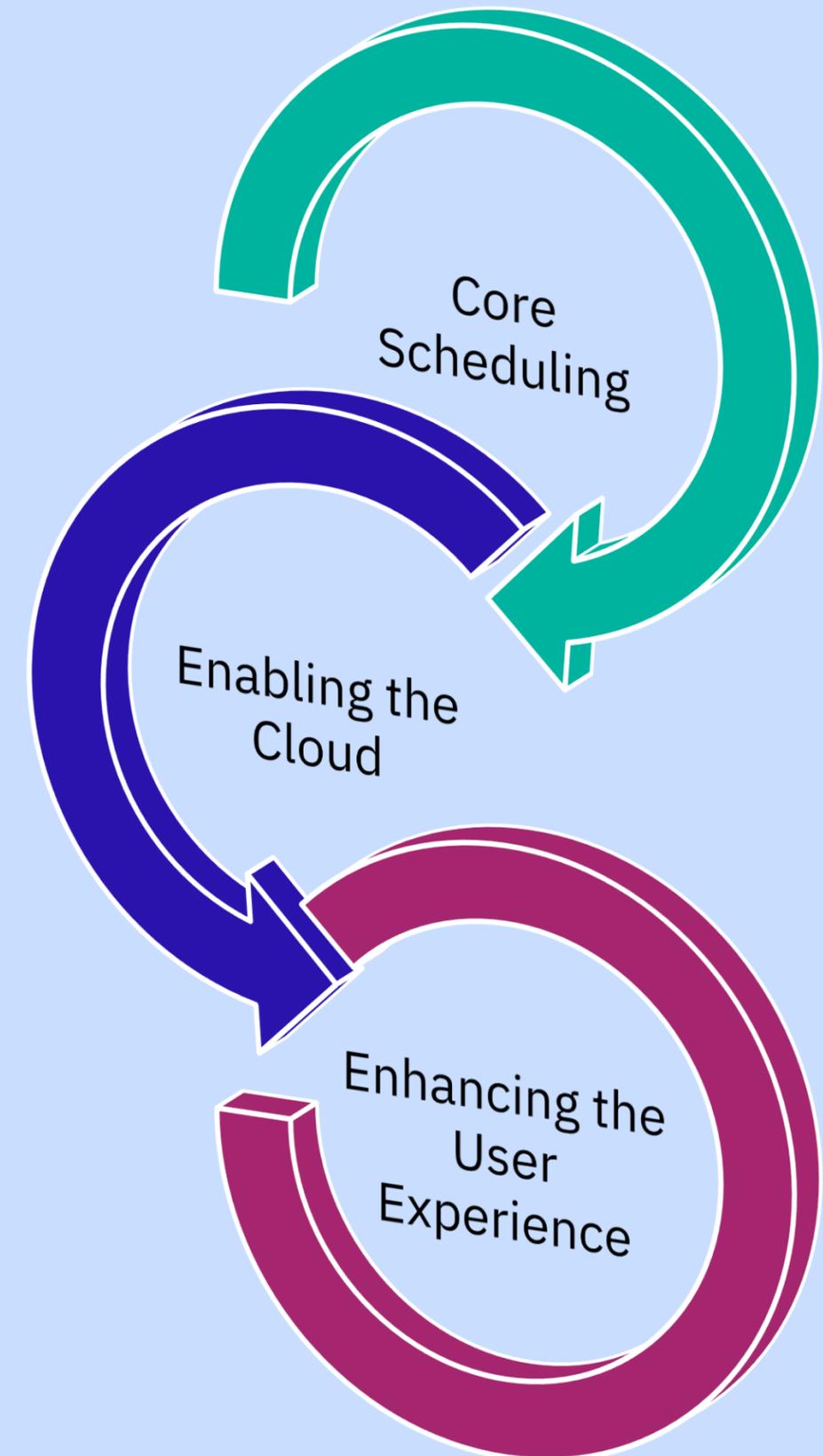
- **Performance & Scalability**
- **Workload Policies**
- **Technology: GPU's & Containers**

Enabling Multi-Cloud

- **When to forward work**
- **Ensuring the right data is available**
- **Intelligent autoscaling the Cloud**

Enhancing the User Experience

- **Simplifying HPC**
- **Computational Workflows**
- **Operational Visibility**



Cutting Edge Performance and Scalability

Continuous Performance Improvement:

LSF 10 delivers ~3.8x improvement in scheduling performance.

Every update contains new performance optimizations – 6 monthly updates.

Scalability:

The largest single cluster today is in excess of 12,000 hosts running large scale parallel simulations.

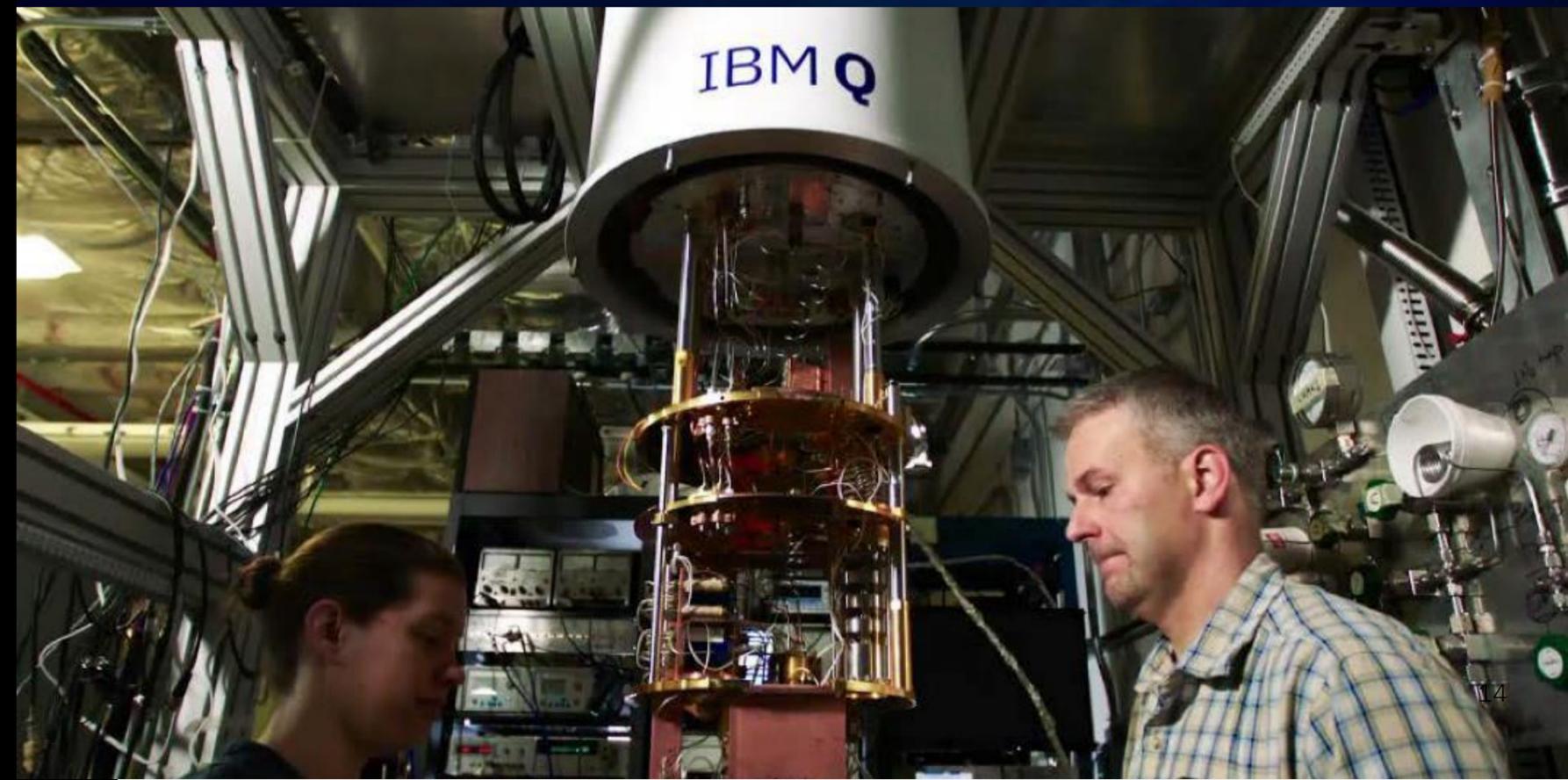
The largest high throughput cluster is around 6,500 hosts, running 7M+ jobs per day, and bursting 50,000+ cores to the cloud

CORAL and IBM Q

Sampling of OpenPOWER Members Contributing to Sierra & Summit



2/3^{rds} of these use LSF

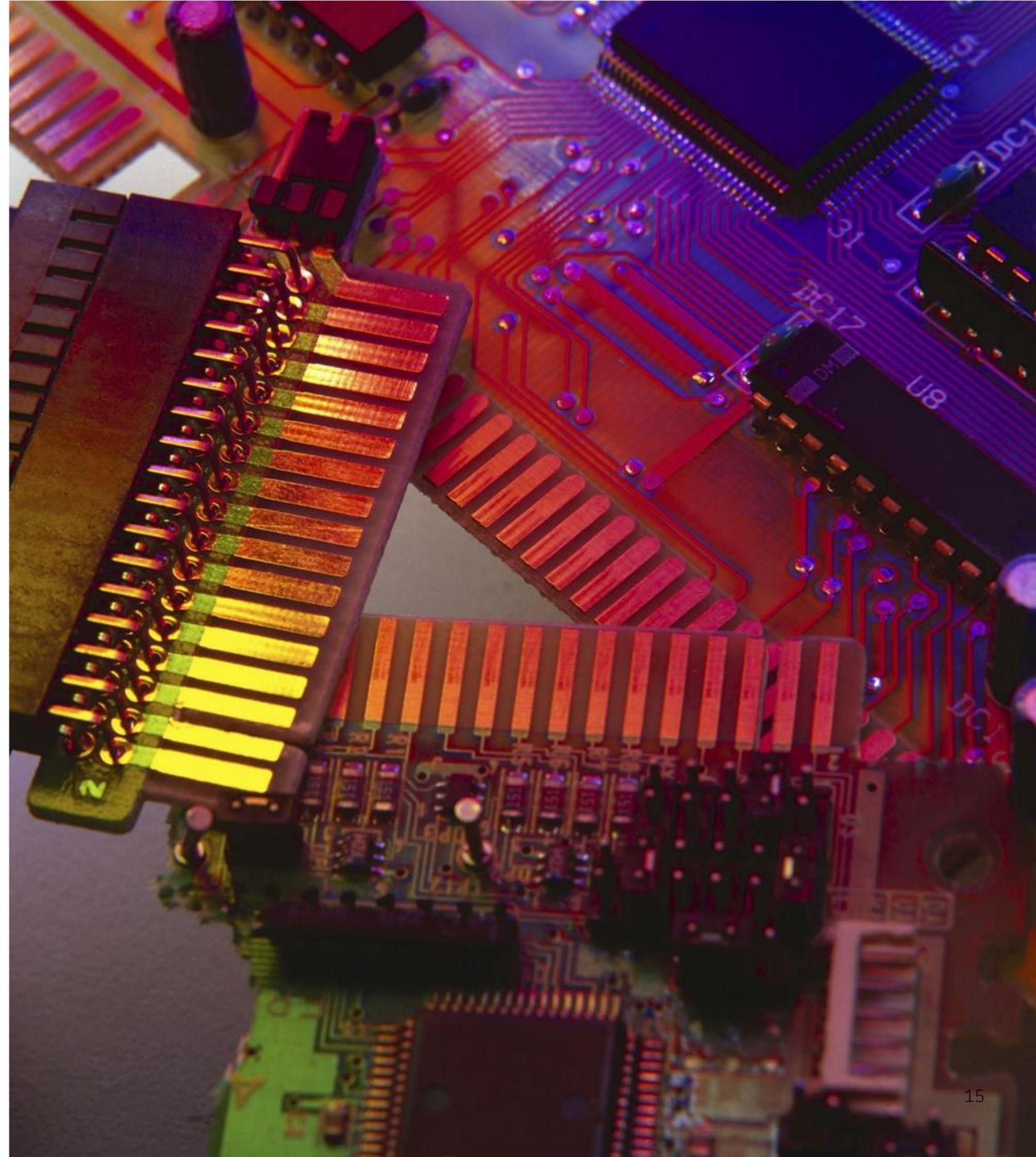


Market Leading GPU Support

Over 10 years of GPU functionality

Recent enhancements:

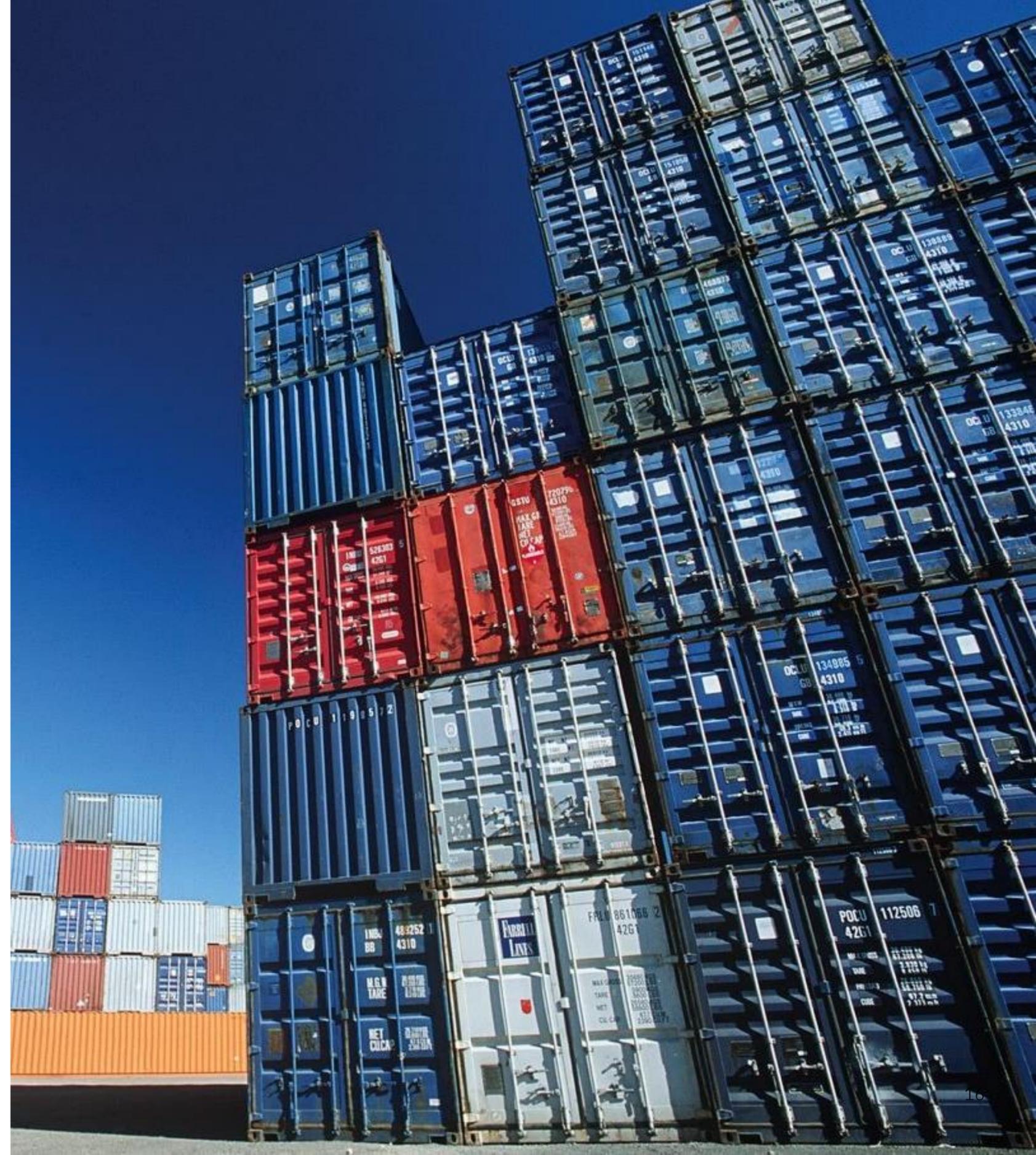
- “Zero Config” - LSF will now automatically detect and configure GPU support. This means that users can take advantage of GPU's as soon as Spectrum LSF installed.
- Simplified (-gpu) syntax
- GPU Fairshare & GPU Pre-emption
- Multi-MPS Support – Multiple MPS daemons per job and/or multiple jobs per MPS daemon (spk8)
- Additional affinity options (spk8)



Running Containerized Workloads with LSF

LSF 10 provides:

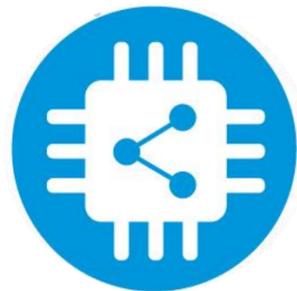
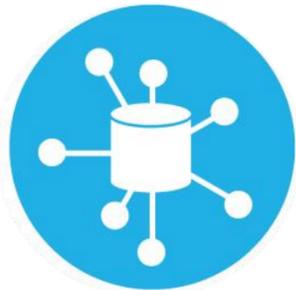
- Integrated support for running containerized applications with Docker, Shifter and Singularity.
- Transparent container access – users don't need to learn complex container syntax.
- All container startup and filesystem mounting is performed by LSF.
- The User never gains elevated privileges.
- Administrator Visibility of container use:
 - host, container name, tags, source repository, file path, size, install time, age, last used, last used by



HPC Administration in the Cloud Era

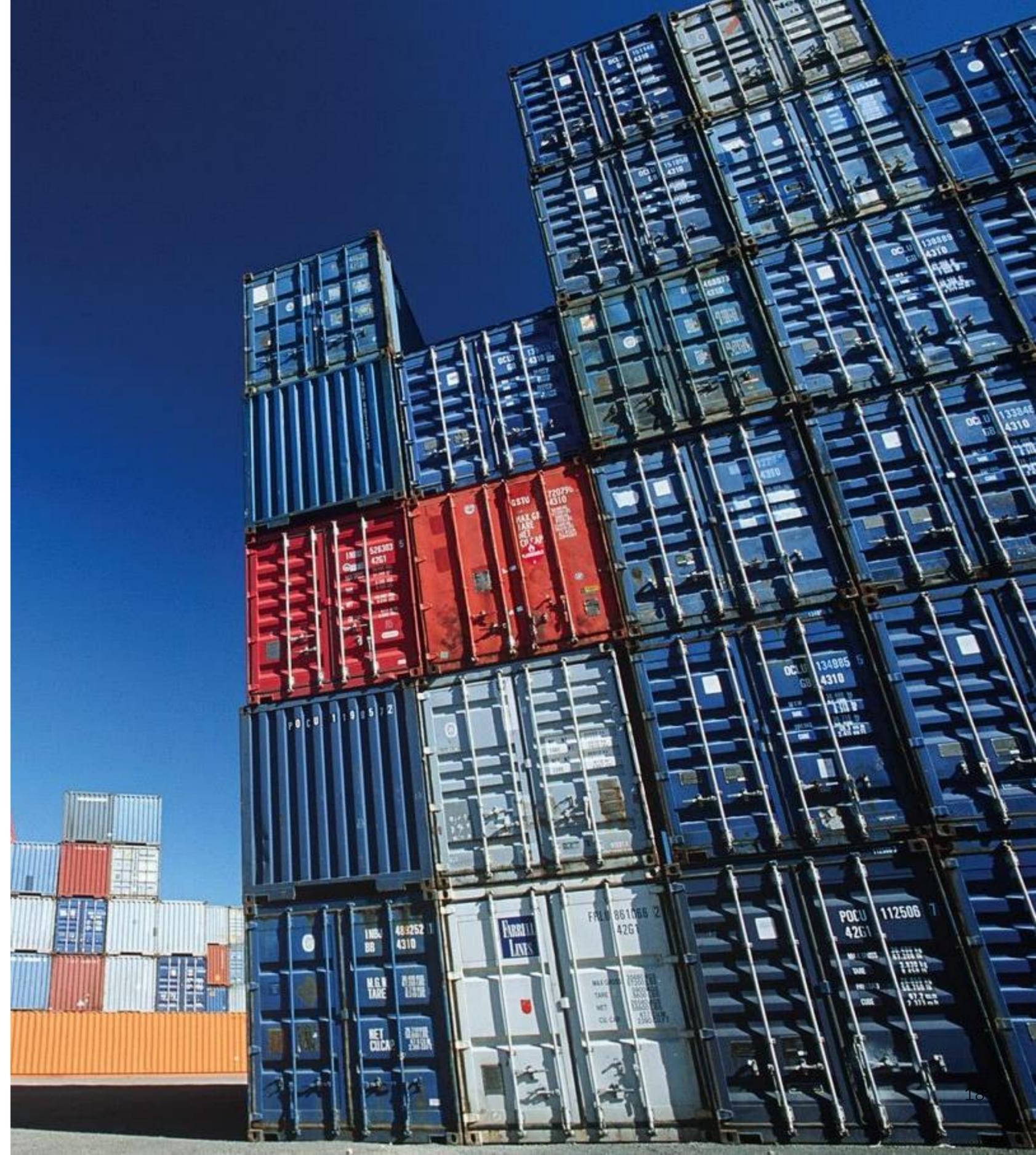
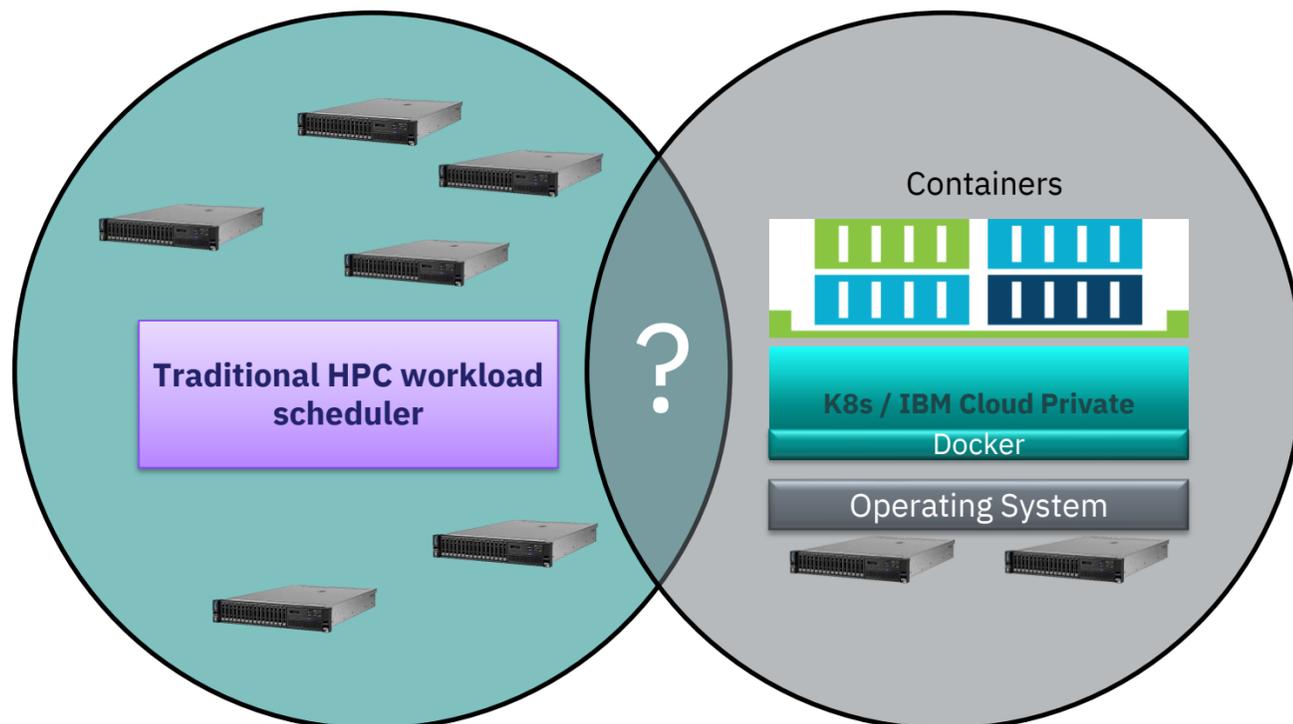
Numerous considerations and requirements to satisfy

- I need to on-board cloud native workloads without disrupting my existing HPC cluster
- I need enable DevOps tools and processes for my users
- I need to provide a secure multi-user environment that doesn't sacrifice performance



HPC and Kubernetes

- Can k8s address all HPC use cases and negate the need for a traditional workload scheduler?
- While many high performance computing workloads can be containerized, they expect various services to be available in the environment – containerizing these can be challenging.
- How to integrate k8s with a traditional workload scheduler to get the best of both worlds?

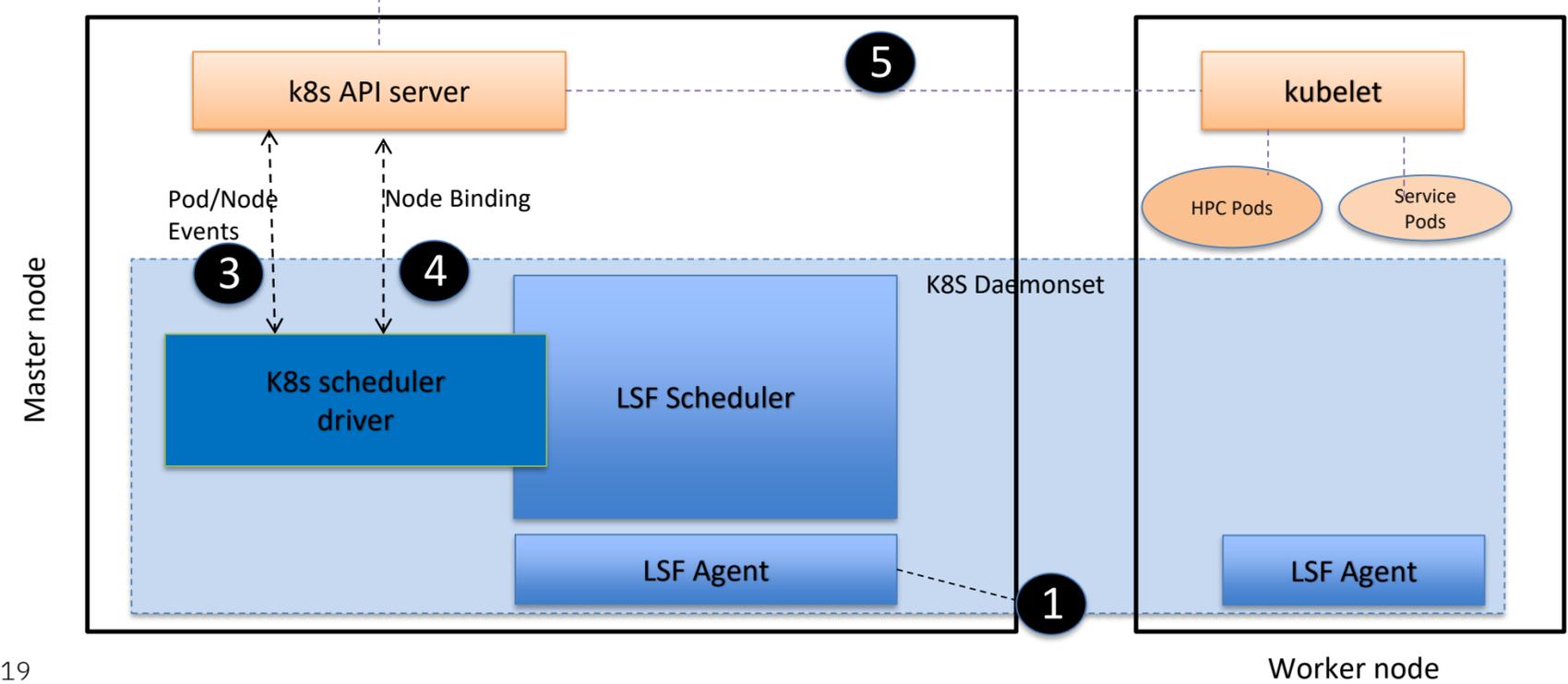


Deployment Option #1: LSF as a pod scheduler in a pure K8S environment

```
apiVersion: v1
kind: Pod
metadata:
  name: myapp-pod
  labels:
    app: myapp
  annotations:
    lsf.ibm.com/queue: "night"
    lsf.ibm.com/fairshareGroup: "project-1"
    lsf.ibm.com/gpu: "gpu=4:mode=shared"
spec:
  schedulerName: lsf
  containers:
  - name: myapp-container
    image: busybox
    command: ['sh', '-c', 'echo Hello
Kubernetes! && sleep 3600']
```

2

kubectl



K8S native user experience

1. HPC Scheduler components are deployed as DaemonSet in K8S via Helm chart
2. Users submit workload into K8S API annotating pods with scheduler directives
3. Driver listens to API servers and translates pod requests into jobs in HPC Scheduler
4. HPC Scheduler makes decisions to bind pod to specific node based on policy
5. Kubelet executes and manages pod lifecycle on target nodes

Deployment Option #2: Existing Spectrum LSF augmented with Kubernetes

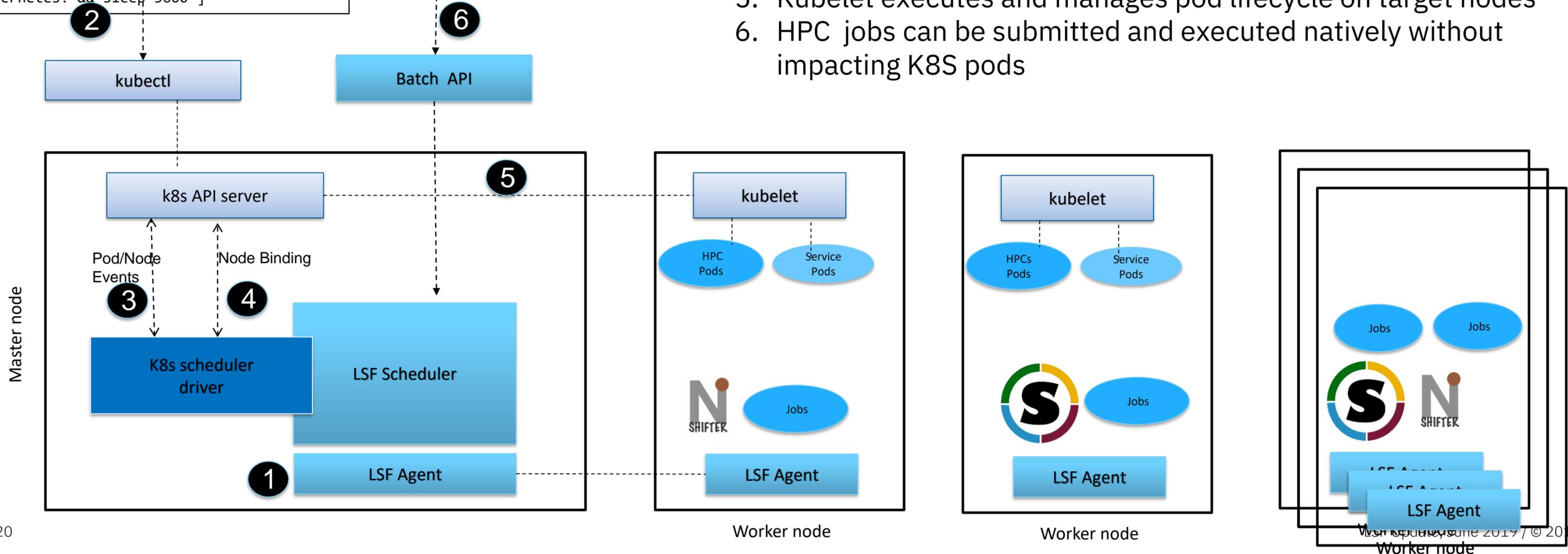
```

apiVersion: v1
kind: Pod
metadata:
  name: myapp-pod
  labels:
    app: myapp
  annotations:
    lsf.queue: "night"
    lsf.fairshareGroup: "project-1"
    lsf.ibm.com/gpu: "gpu=4:mode=shared"
spec:
  schedulerName: lsf
  containers:
    - name: myapp-container
      image: busybox
      command: ['sh', '-c', 'echo Hello
Kubernetes! && sleep 3600']
  
```

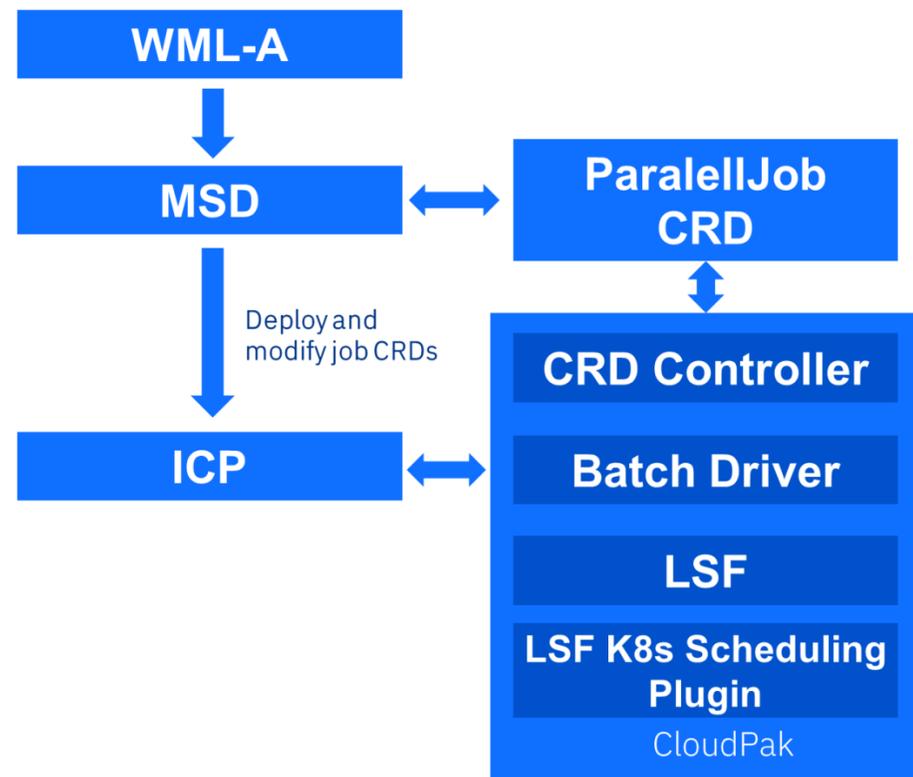
```

bsub -app singularity -n 5,10
mpijob.sh
  
```

1. LSF runs on bare OS, and connects to a k8s API server. LSF cluster can be larger than K8S nodes (>5000 nodes)
2. Users submit workload into K8S API annotating pods with scheduler directives
3. Driver listens to API servers and translates pod requests into jobs in LSF Scheduler
4. LSF Scheduler makes decisions to bind pod to specific node based on policy
5. Kubelet executes and manages pod lifecycle on target nodes
6. HPC jobs can be submitted and executed natively without impacting K8S pods



Watson Machine Learning Accelerator Use Case



```

apiVersion: scheduling.batch/v1alpha1
kind: ParallelJob
metadata:
  name: large_model_train
  namespace: default
  annotations:
    lsf.ibm.com/queue: myQueue
spec:
  name: large_model_train #same with metadata/name
  description: This is a parallel job to run model training across hosts.
  priority: 100
  headerTask: group0
  placement:
    sameTerm: Zone | Rack | Host
  taskGroups:
  - metadata:
    name: parameter_server
    spec:
      replica: 1
      template:
        spec:
          containers:
            image: ubuntu
            resources:
              request:
                cpu: 1
                memory: 4096Mi
  
```

← LSF specific annotations

← Job level specification

← Task level specification

```

- metadata:
  name: worker_nodes
  annotations:
    lsf.ibm.com/gpu: nvlink=yes
spec:
  placement:
    spanTerm:
      - topologyKey: node
      taskTile: 2
      - topologyKey: node
      taskTile: 4
  replica: 4
  template:
    spec:
      containers:
        image: nvida/cuda9.2
        name: task1
        resources:
          request:
            nvidia.com/gpu: 2
            memory: 16000Mi
  
```



IBM
Spectrum
LSF

http://ibm.biz/SpectrumComputing_CloudPak_preview

IBM



IBM
Spectrum
Computing

Management visibility LSF Explorer & LSF RTM

The screenshot displays the LSF Explorer interface with the following components:

- Header:** IBM Spectrum LSF Suite for HPC 10.2.0.6, navigation tabs for Workload, Resources, System & Settings, and Reports, user profile (Isfadmin), and system time (8:38:23 PM +0000).
- Left Sidebar:** Navigation icons for Home, Refresh, and Back.
- Main Area:** A grid of 8 dashboard widgets:
 - Charge Back:** Shows workload size change distribution and a summary table.
 - Cluster Overview:** Provides a high-level view of cluster resources and usage.
 - FlexLM License Usage:** Displays peak and average license usage by user and a summary table.
 - Hardware:** Monitors cluster availability and hardware utilization.
 - Pending Job Analysis:** Analyzes pending jobs with a bar chart and summary table.
 - Workload Accounting:** Shows job run time and job peak time statistics.
 - Workload Scale IO Accounting:** Visualizes workload scale and IO accounting data.
 - Workload Statistics:** Presents active workload statistics and a summary table.
- Right Side:** '+ New' and 'Delete' buttons for dashboard management.

Explorer enhancements for SPK8/SPK9

- Update to ElasticSearch 6.x
- Performance Optimizations
- Data consistency between LSF CLI and ES

RTM 10.2 enhancements:

- New responsive user interface with an updated look and feel that is consistent with the rest of the LSF family GUI's.
- Simplified installation
- Enhanced performance with Cacti 1.2