HTCondor and TORQUE/Maui

Resource Management Integration

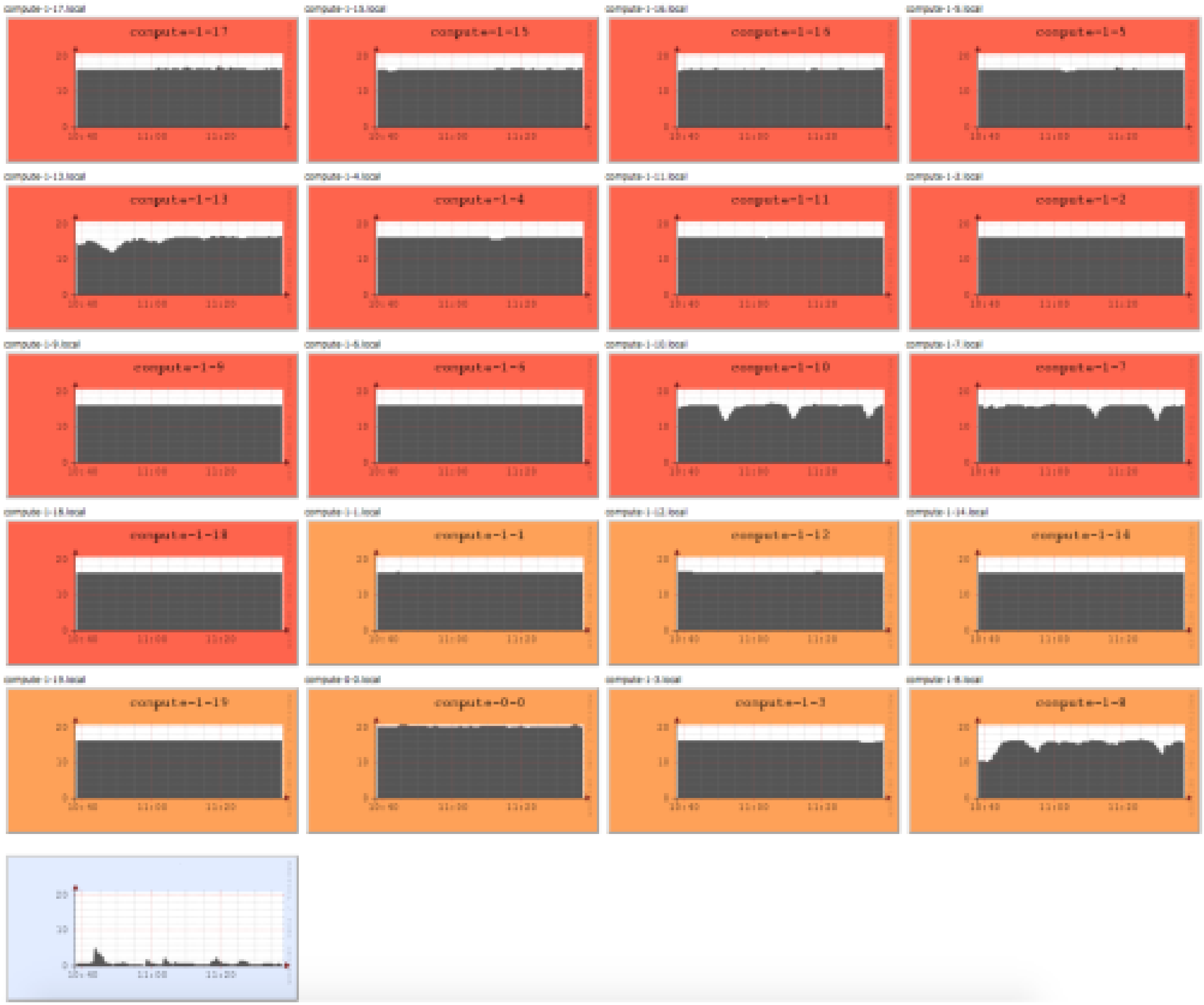
Grigoriy Abramov

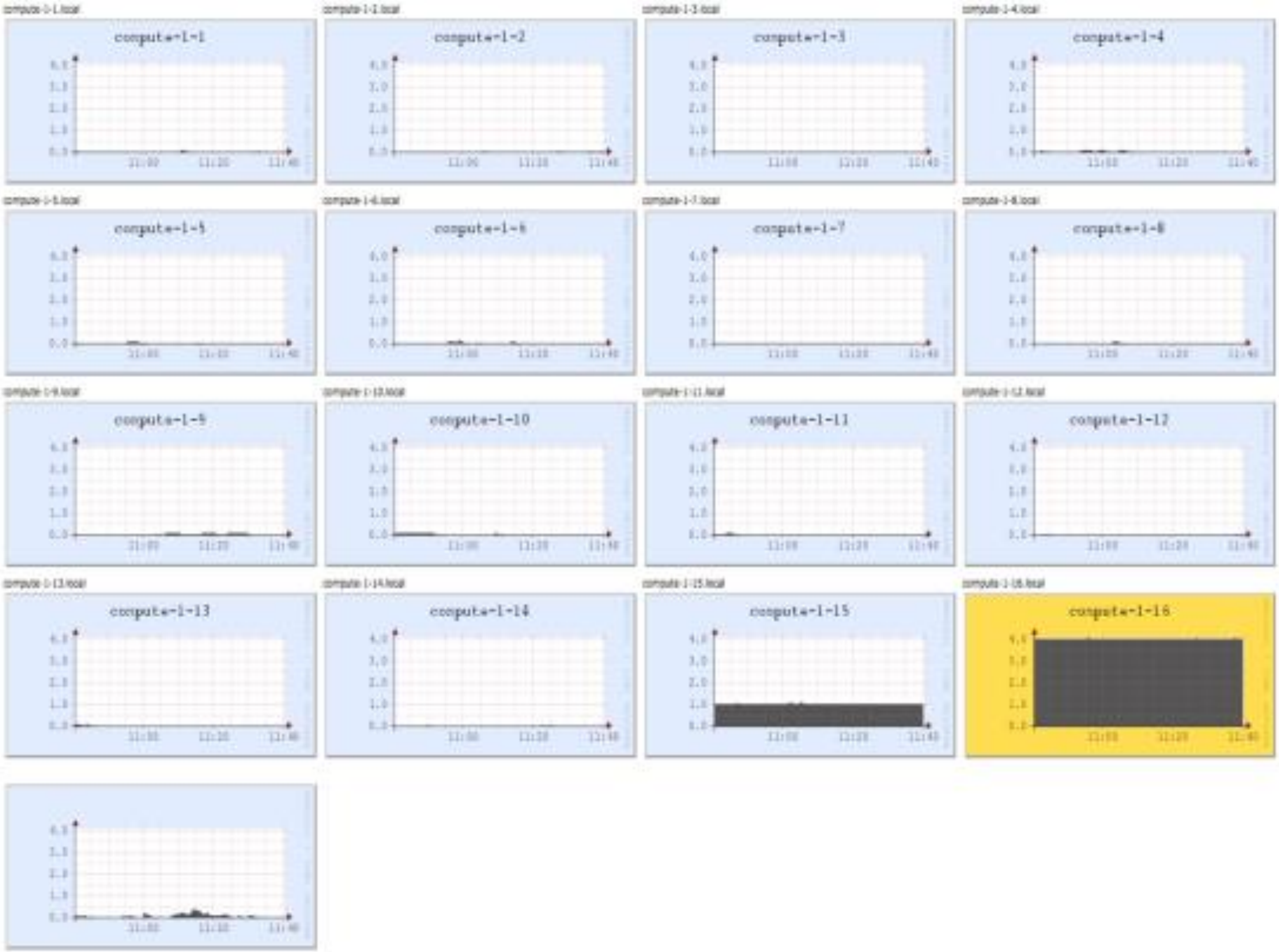
Systems Administrator, Research Data Center

Chicago, IL

Problems:

- Low efficiency of available computational resources
- Disproportionate utilization of each individual cluster
- Absence of sharing computational resources
- Absence of a unified, resource-management platform
- Variations on Linux OS releases
- Distributed ownership of computational resources


compute-1-17.local
compute-1-17
compute-1-15.local
compute-1-15
compute-1-16.local
compute-1-16
compute-1-5.local
compute-1-5
compute-1-13.local
compute-1-13
compute-1-4.local
compute-1-4
compute-1-11.local
compute-1-11
compute-1-2.local
compute-1-2
compute-1-9.local
compute-1-9
compute-1-6.local
compute-1-6
compute-1-10.local
compute-1-10
compute-1-7.local
compute-1-7
compute-1-18.local
compute-1-18
compute-1-1.local
compute-1-1
compute-1-12.local
compute-1-12
compute-1-14.local
compute-1-14
compute-1-19.local
compute-1-19
compute-0-0.local
compute-0-0
compute-1-3.local
compute-1-3
compute-1-8.local
compute-1-8
20
10
0
10:40
11:00
11:20


compute-1-1.local
compute-1-1
compute-1-2.local
compute-1-2
compute-1-3.local
compute-1-3
compute-1-4.local
compute-1-4
compute-1-5.local
compute-1-5
compute-1-6.local
compute-1-6
compute-1-7.local
compute-1-7
compute-1-8.local
compute-1-8
compute-1-9.local
compute-1-9
compute-1-10.local
compute-1-10
compute-1-11.local
compute-1-11
compute-1-12.local
compute-1-12
compute-1-13.local
compute-1-13
compute-1-14.local
compute-1-14
compute-1-15.local
compute-1-15
compute-1-16.local
compute-1-16
4.0
3.0
2.0
1.0
0.0
11:00
11:20
11:40

Advantage:

Availability of TORQUE/Maui resource management
and Portable Batch System (PBS)
on all HPC computational resources

Goals:

- Cost-efficient optimization of computational resources
- Greater access to computational resources for faculty members and research groups who do not have the resources to obtain their own clusters

Working Principles:

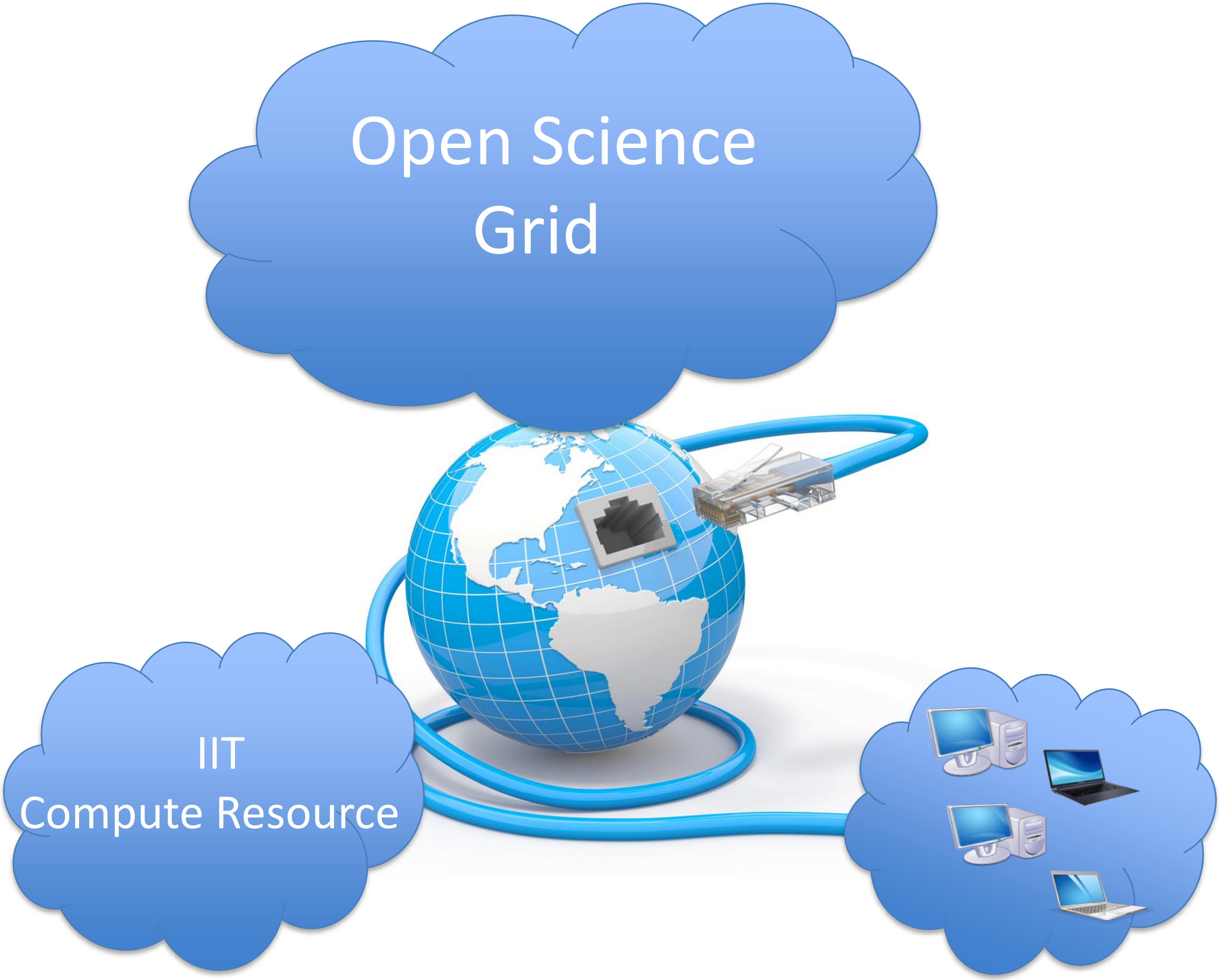
- Departments and research groups retain full ownership of their clusters and have priority in performing their computations
    - All running computational jobs submitted by "guests" on shared resources should be removed from the queue when needed

Challenges:

- Finding an optimal platform that does not require system reinstallation or significant, configuration updates that would interrupt an already-running computation

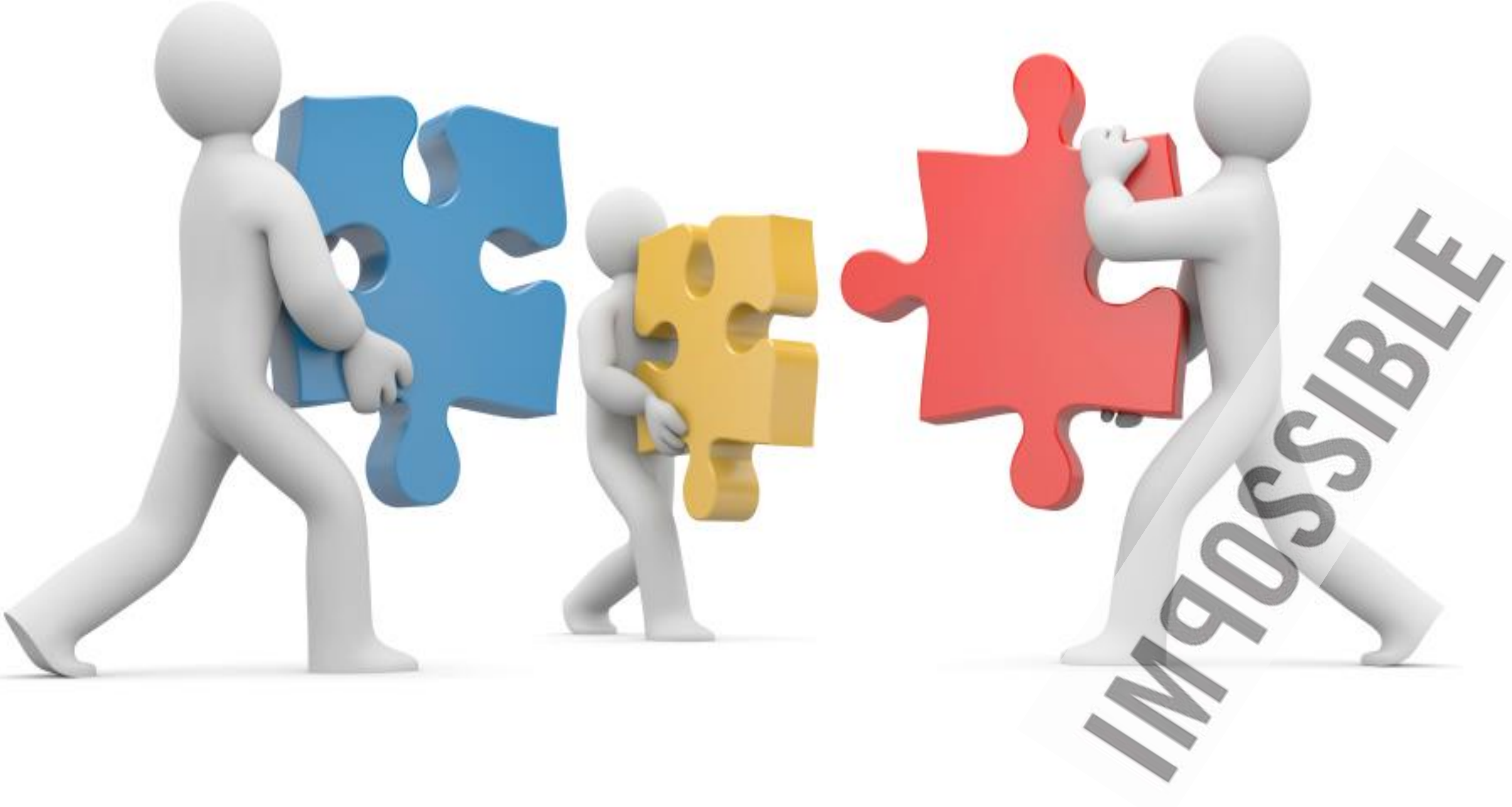
OS and Applications:

- ROCKS cluster OS
- TORQUE/Maui (Adaptive Computing)
- HTCondor, HTCondor-CE-BOSCO, or both

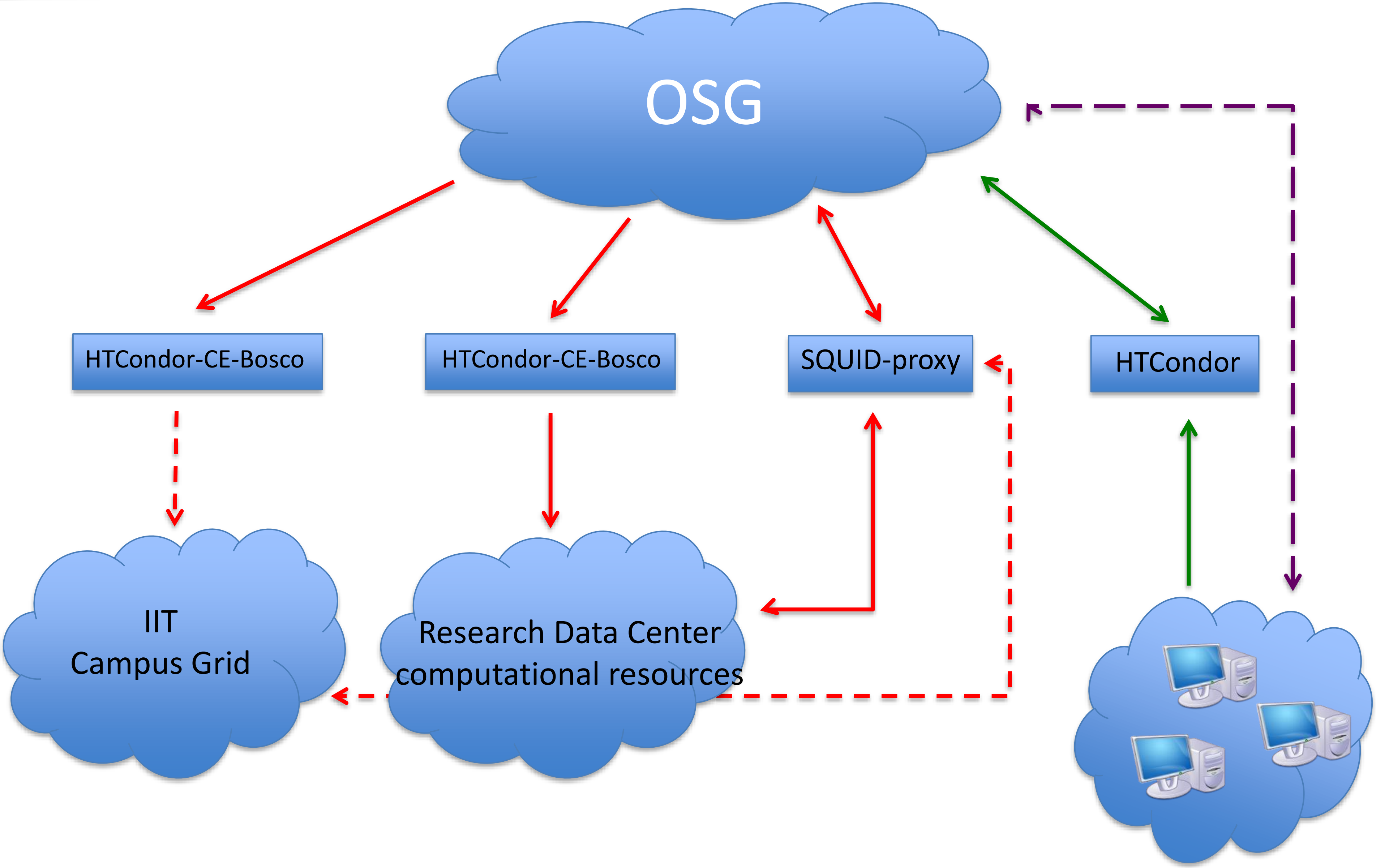


Open Science Grid

IIT Compute Resource


IMPOSSIBLE

GridIIT/OSG Computing Grid Diagram



OSG

HTCondor-CE-Bosco

HTCondor-CE-Bosco

SQUID-proxy

HTCondor

IIT
Campus Grid

Research Data Center
computational resources

Implementation:

- Single-user account and GID on all HPC clusters:
  guestuser/guestuser
- PBS Queue-Manager (qmgr) configuration
- Maui configuration
- Installation of HTCondor-CE-Bosco or HTCondor
- Access to computational resources via Secure Shell (SSH)
- Testing computing grid resources for incoming/outgoing traffic

Implementation | TORQUE configuration

Create and define queue grid at qmgr prompt:

```
create queue grid
set queue grid queue_type = Execution
set queue grid max_user_queuable = 14
set queue grid resources_default.walltime = 48:00:00
set queue grid resources_default.ncpus = 1
set queue grid acl_group_enable = True
set queue grid acl_groups = guestuser
set queue grid kill_delay = 120
set queue grid keep_completed = 120
set queue grid enabled = True
set queue grid started = True
```

Implementation | Maui configuration

- Priority
- Preemption
- Preemption policy
- Partitioning
- QOS – Quality of Services

Implementation | Maui configuration (cont.)

```
RMCFG[base]  TYPE=PBS          SUSPENDSIG=15
PREEMPTPOLICY                  SUSPEND
NODEALLOCATIONPOLICY           PRIORITY
QOSCFG[hi]                     QFLAGS=PREEMPTOR
QOSCFG[low]                    QFLAGS=NOBF:PREEMPTEE
CLASSWEIGHT                    10
CLASSCFG[batch]                QDEF=hi   PRIORITY=1000
CLASSCFG[grid]                 QDEF=low PRIORITY=1
```

Implementation | MAUI configuration (cont.)

```
GROUPCFG[users]       PRIORITY=1000   QLIST=hi    QDEF=hi    QFLAGS=PREEMPTOR
GROUPCFG[guestuser]   PRIORITY=1      QLIST=low   QDEF=low   QFLAGS=PREEMPTEE
USERCFG[guestuser]    PRIORITY=1      QLIST=low   QDEF=low   QFLAGS=PREEMPTEE

PARTITIONMODE ON
NODECFG[compute-1-1]     PARTITION=grid
NODECFG[compute-1-2]     PARTITION=grid

SYSCFG[base]            PLIST=default, grid&
USERCFG[DEFAULT]        PLIST=default
GROUPCFG[guestuser]     PLIST=default:grid   PDEF=default
```

* Maui service needs to be restarted

- Test job submission via PBS as guestuser on compute cluster
- In submit script, the below-listed options should be presented:

```
#PBS -q grid
#PBS -W x=“PARTITION:grid”
```

- Reliability of PREEMPTION needs to be verified
- Install and configure HTCondor or HTCondor-CE-BOSCO
- Add on computational cluster’s head node following lines to file

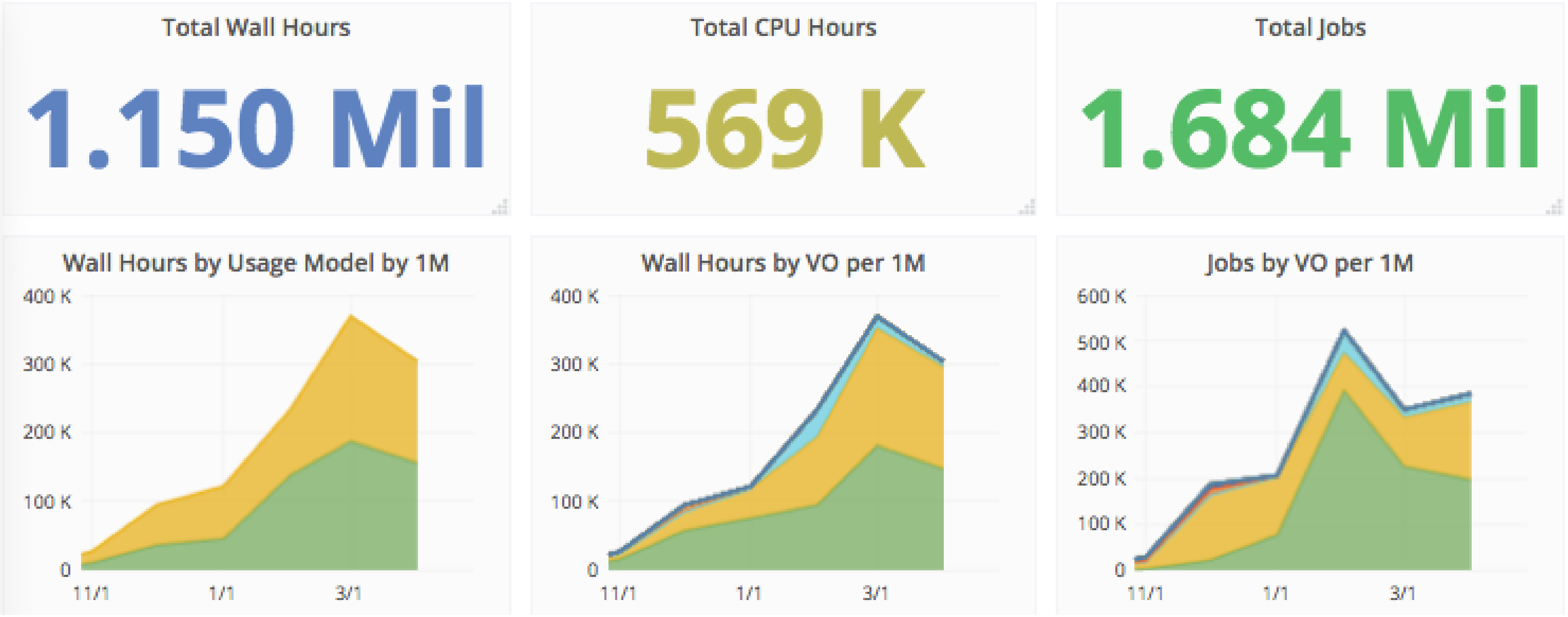
../bosco/glite/bin/pbs_local_submit_attributes.sh

```
#!/bin/sh
echo "#PBS -q grid”
echo  '#PBS -W x="PARTITION:grid”’
```

- Submit test job from remote server via command:

```
bosco_cluster –t  guestuser@your_cluster_name.edu
```

GridIIT and OSG Shared Computational Resources Over a 6-Month Period



Total Wall Hours

1.150 Mil

Total CPU Hours

569 K

Total Jobs

1.684 Mil

*Opportunistic *Dedicated

Source: *http://gracc.opensciencegrid.org*