Pegasus

Enhancing User Experience on OSG

Mats Rynge
rynge@isi.edu

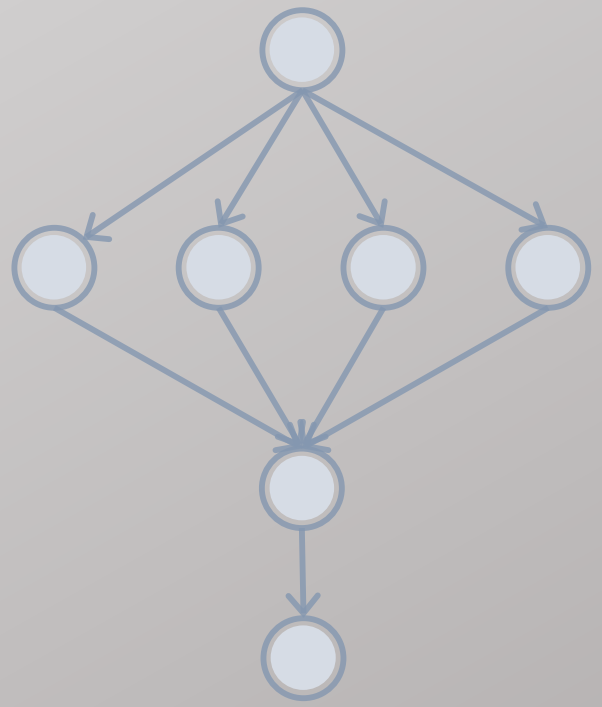
USC Viterbi
School of Engineering
Information Sciences Institute

https://pegasus.isi.edu

Key Pegasus Concepts

Pegasus WMS == Pegasus planner (mapper) + DAGMan workflow engine + HTCondor scheduler/broker

- Pegasus maps workflows to infrastructure
- DAGMan manages dependencies and reliability
- HTCondor is used as a broker to interface with different schedulers



Workflows are DAGs (or hierarchical DAGs)

- Nodes: jobs, edges: dependencies
- No while loops, no conditional branches

Planning occurs ahead of execution

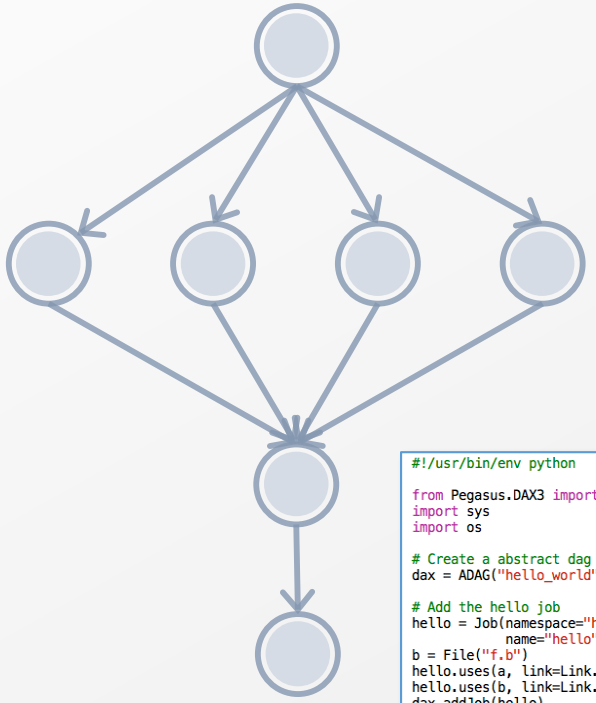
- (Except hierarchical workflows)

Planning converts an abstract workflow into a concrete, executable workflow

- Planner is like a compiler

DAX

DAG in XML



```python
#!/usr/bin/env python

from Pegasus.DAX3 import *
import sys
import os

# Create a abstract dag
dax = ADAG("hello_world")

# Add the hello job
hello = Job(namespace="hello_world",
            name="hello", version="1.0")
b = File("f.b")
hello.uses(a, link=Link.INPUT)
hello.uses(b, link=Link.OUTPUT)
dax.addJob(hello)

# Add the world job (depends on the hello job)
world = Job(namespace="hello_world",
            name="world", version="1.0")
c = File("f.c")
world.uses(b, link=Link.INPUT)
world.uses(c, link=Link.OUTPUT)
dax.addJob(world)

# Add control-flow dependencies
dax.addDependency(Dependency(parent=hello,
                             child=world))

# Write the DAX to stdout
dax.writeXML(sys.stdout)
```

DAG

directed-acyclic graphs

stage-in job

Transfers the workflow input data

clustered job

Groups small jobs together
to improve performance

cleanup job

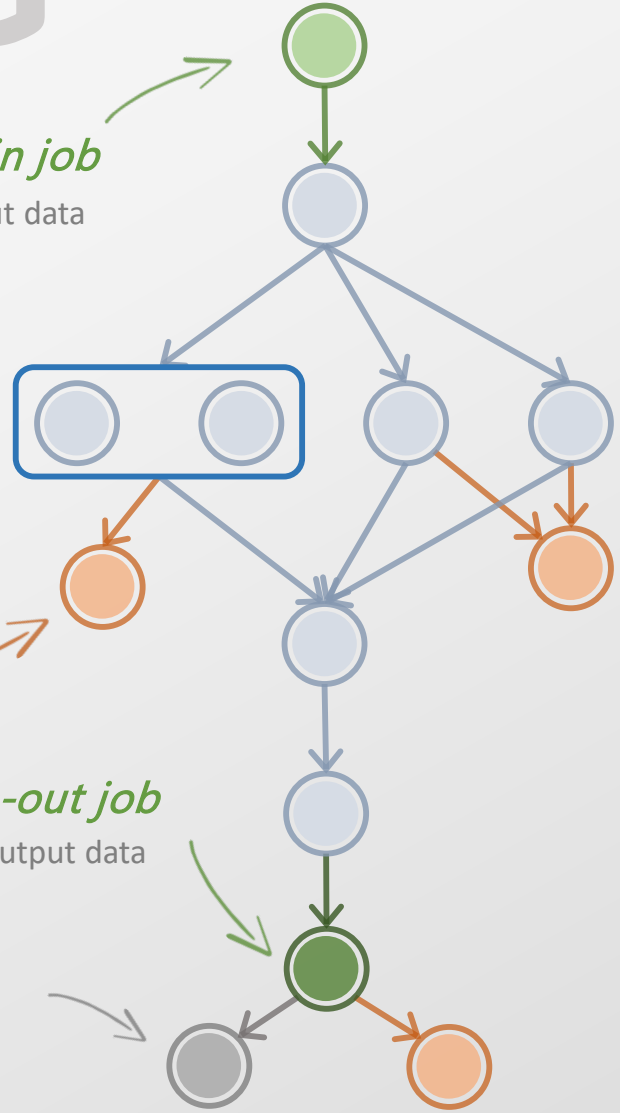
Removes unused data

stage-out job

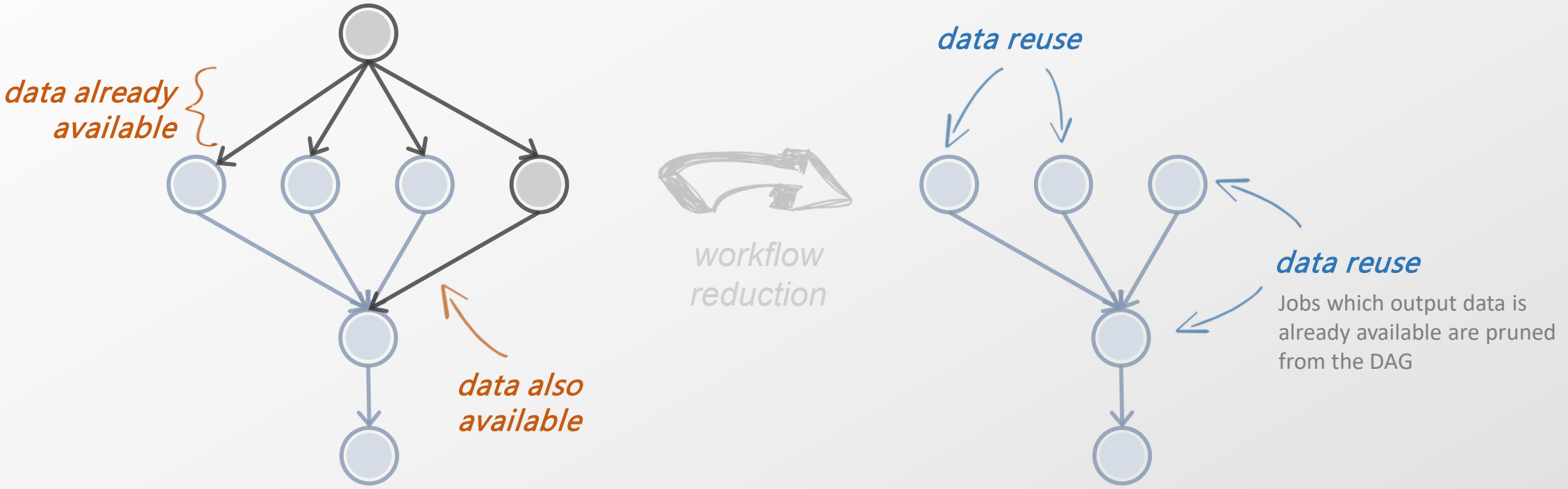
Transfers the workflow output data

registration job

Registers the workflow output data



What about **data reuse**?



Data Staging Configurations

- Condor I/O (HTCondor pools, OSG, …)
  - Worker nodes do not share a file system
  - Data is pulled from / pushed to the submit host via HTCondor file transfers
  - Staging site is the submit host
- Non-shared File System (clouds, OSG, …)
  - Worker nodes do not share a file system
  - Data is pulled / pushed from a staging site, possibly not co-located with the computation
- Shared File System (HPC sites, XSEDE, Campus clusters, …)
  - I/O is directly against the shared file system

pegasus-transfer

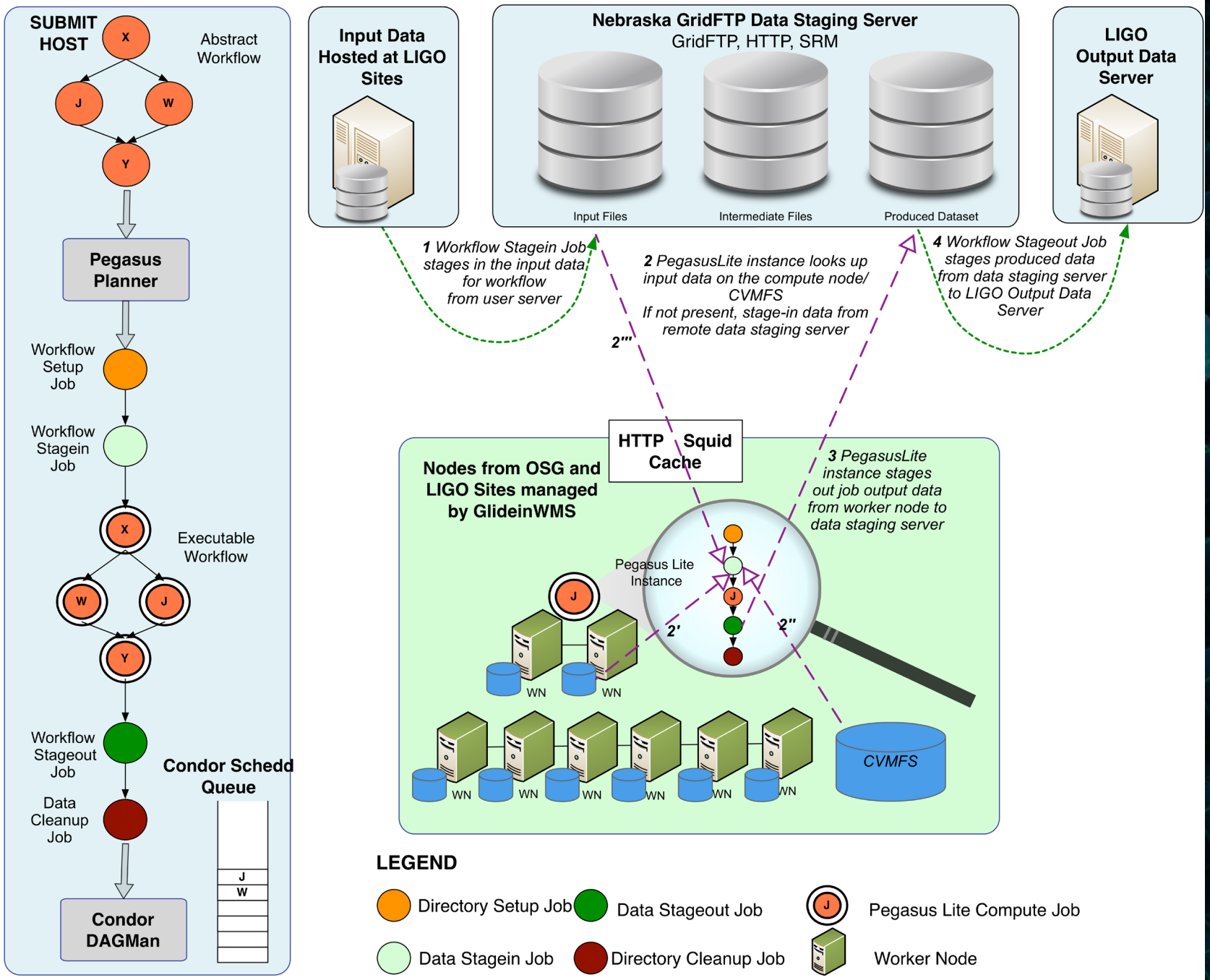
- Pegasus' internal data transfer tool with support for a number of different protocols
- Directory creation, file removal
  - If protocol supports, used for cleanup
- Two stage transfers
  - e.g. GridFTP to S3 = GridFTP to local file, local file to S3
- Parallel transfers
- Automatic retries
- Credential management
  - Uses the appropriate credential for each site and each protocol (even 3rd party transfers)

HTTP
SCP
GridFTP
Globus Online
iRods
Amazon S3
Google Storage
SRM
FDT
stashcp
cp
ln -s

$OSG_SQUID_LOCATION / http_proxy

- $OSG_SQUID_LOCATION is set by many sites
  - But does it work?
  - Does it work for the particular http source the user needs?
- pegasus-transfer will use $OSG_SQUID_LOCATION if
  - http_proxy is not specified by the user
  - for the first transfer attempt

Data Flow for LIGO Pegasus Workflows in OSG



Advanced LIGO – Laser Interferometer Gravitational Wave Observatory

Replica catalog – multiple sources

pegasus.conf

```
# Add Replica selection options so that it will try URLs first, then
# XrootD for OSG, then gridftp, then anything else
pegasus.selector.replica=Regex
pegasus.selector.replica.regex.rank.1=file:///cvmfs/.*
pegasus.selector.replica.regex.rank.2=file://.*
pegasus.selector.replica.regex.rank.3=root://.*
pegasus.selector.replica.regex.rank.4=gridftp://.*
pegasus.selector.replica.regex.rank.5=.\*
```

Replica Catalog

```
# This is the replica catalog. It lists information about each of the
# input files used by the workflow. You can use this to specify locations
# to input files present on external servers.

# The format is:
# LFN PFN site="SITE"

f.a   file:///cvmfs/oasis.opensciencegrid.org/diamond/input/f.a   site="cvmfs"
f.a   file:///local-storage/diamond/input/f.a   site="prestaged"
f.a   gridftp://storage.mysite/edu/examples/diamond/input/f.a   site="storage"
```

pegasus-kickstart

() login02.osgconnect.net — Konsole

File Edit View Bookmarks Settings Help


```
<?xml version="1.0" encoding="UTF-8"?>
<invocation xmlns="http://pegasus.isi.edu/schema/invocation" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://pegasus.is
i.edu/schema/invocation http://pegasus.isi.edu/schema/iv-2.3.xsd" version="2.3" start="2016-11-28T14:27:48.909-06:00" duration="11200.691" transformatio
n="job-wrapper.sh" derivation="ID0013214" resource="condorpool" wf-label="particleshower" wf-stamp="2016-11-22T21:14:13-06:00" interface="eth0" hostaddr
="131.225.208.240" hostname="fnpc4593.fnal.gov" pid="1725084" uid="12740" user="osg" gid="9652" group="osg" umask="0022">
  <mainjob start="2016-11-28T14:27:49.007-06:00" duration="11200.593" pid="1725089">
    <usage utime="10921.591" stime="30.304" maxrss="395820" minflt="128741" majflt="18" nswap="0" inblock="85776" outblock="1717424" msgsnd="0" msgrcv="
0" nsignals="0" nvcsw="7676" nivcsw="185495"/>
    <status raw="0"><regular exitcode="0"/></status>
    <statcall error="0">
      <file name="/storage/local/data1/condor/execute/dir_1227464/glide_bSxwfe/execute/dir_1724937/pegasus.XRZ1p3/job-wrapper.sh">23212F62696E2F62617368
0A0A736574</file>
      <statinfo mode="0100755" size="1305" inode="16648869" nlink="1" blksize="4096" blocks="8" mtime="2016-11-28T12:10:53-06:00" atime="2016-11-28T14:2
7:48-06:00" ctime="2016-11-28T14:27:48-06:00" uid="12740" user="osg" gid="9652" group="osg"/>
    </statcall>
    <argument-vector>
      <arg nr="1">100</arg>
      <arg nr="2">0</arg>
      <arg nr="3">gamma</arg>
      <arg nr="4">62</arg>
      <arg nr="5">VERITAS</arg>
      <arg nr="6">corsika.tar.gz</arg>
      <arg nr="7">corsika75000Linux_QGSII_urqmd</arg>
      <arg nr="8">13213
    </argument-vector>
  </mainjob>
  <jobids condor="547839.0"/>
  <cwd>/storage/local/data1/condor/execute/dir_1227464/glide_bSxwfe/execute/dir_1724937/pegasus.XRZ1p3</cwd>
  <usage utime="0.013" stime="0.085" maxrss="828" minflt="2448" majflt="0" nswap="0" inblock="0" outblock="0" msgsnd="0" msgrcv="0" nsignals="0" nvcsw="
1" nivcsw="12"/>
  <machine page-size="4096">
    <stamp>2016-11-28T14:27:48.909-06:00</stamp>
    <uname system="linux" nodename="fnpc4593.fnal.gov" release="2.6.32-642.6.2.el6.x86_64" machine="x86_64">#1 SMP Tue Oct 25 15:06:33 CDT 2016</uname>
    <linux>
      <ram total="65319608" free="1071948" shared="0" buffer="148224"/>
      <swap total="8388604" free="7741364"/>
      <boot idle="45893257.760">2016-11-09T16:40:54.260-06:00</boot>
      <cpu count="32" speed="2000" vendor="AuthenticAMD">AMD Opteron(tm) Processor 6128</cpu>
      <load min1="26.35" min5="27.70" min15="24.33"/>
      <procs total="881" running="23" sleeping="854" waiting="3" zombie="1" vmsize="65009304" rss="14780272"/>
      <task total="1273" running="24" sleeping="1243" waiting="5" zombie="1"/>
    </linux>
```


5,5 Top

() login02.osgconnect.net

```
------------------------------------------------------------------------------
Type           Succeeded Failed  Incomplete  Total     Retries   Total+Retries
Tasks          100000    0       0           100000    543       100543
Jobs           20206     0       0           20206     604       20810
Sub-Workflows  0         0       0           0         0         0
------------------------------------------------------------------------------

Workflow wall time                                       : 19 hrs, 37 mins
Cumulative job wall time                                 : 1 year, 5 days
Cumulative job wall time as seen from submit side        : 1 year, 27 days
Cumulative job badput wall time                          : 2 hrs, 42 mins
Cumulative job badput wall time as seen from submit side : 2 days, 2 hrs
```

```
$ pegasus-analyzer pegasus/examples/split/run0001
pegasus-analyzer: initializing...

****************************Summary
Total jobs : 7 (100.00%)
# jobs succeeded : 7 (100.00%)
# jobs failed : 0 (0.00%)
# jobs unsubmitted : 0 (0.00%)
```

Provenance data can be summarized (*pegasus-statistics*) or used for debugging (*pegasus-analyzer*)

Pegasus est. 2001

Automate, recover, and debug scientific computations.

Get Started

Pegasus Website

http://pegasus.isi.edu

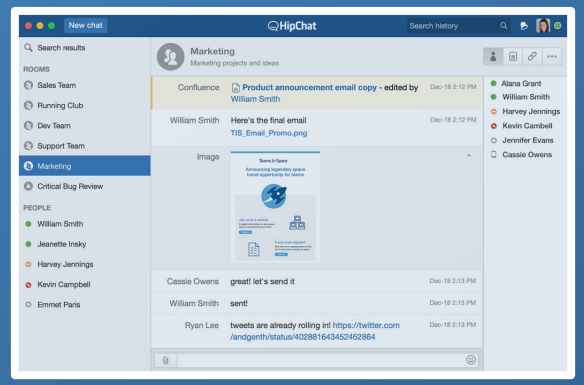
Users Mailing List

pegasus-users@isi.edu

Support

pegasus-support@isi.edu

HipChat



Mats Rynge
rynge@isi.edu

USC Viterbi
School of Engineering
Information Sciences Institute

NSF
U.S. DEPARTMENT OF ENERGY
NIH