



Pegasus WMS Deployments in ACCESS and NAIRR Pilot

Mats Rynge
**University of Southern California
Information Sciences Institute**



Supported by NSF #2138286

APs: Deployments

ACCESS Pegasus

PSC Neocortex/Bridges2

Purdue Anvil Composable
System



EPs: Resource Provisioning

TestPool

HTCondor Annex

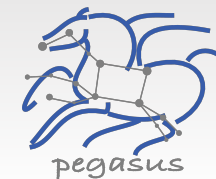
IU Jetstream2 Cloud

pegasus-glidein





Pegasus Workflow Management System



Pegasus WMS

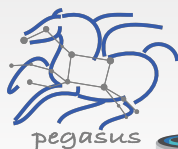
Planner

Monitoring & Provenance

Engine

Scheduler

HTCondor
High Throughput Computing



> Cloud Resources

> OSG

> HPC Systems

> HTCondor Pools

Submit Node

Compute Resources

End to End
Workflow
Management
& Execution

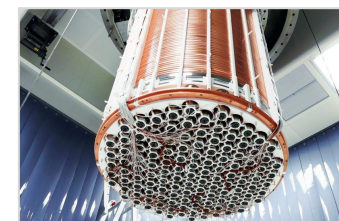
- ▶ Develop portable scientific workflows in Python, Java, and R
- ▶ Compile workflows to be run on heterogeneous resources
- ▶ Monitor and debug workflow execution via CLI and web-based tools
- ▶ Recover from failures with built-in fault tolerance mechanisms
- ▶ Regular release schedule incorporating latest research and development

2001		2003		2005		2007		2009			2011			2013			2015		2017		2018		2020			
		1.0	1.1	1.2	1.3	1.4		2.0	2.1	2.2	2.3	2.4	3.0	3.1	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0	
Development					support for GT4		task clustering		support for AWS			hierarchical workflows		pegasus-lite engine		monitoring dashboard			ensemble manager			support for containers			redesign of APIs	
Research LIGO, SCEC, and others				data cleanup algorithms			data footprint		cloud computing evaluation				MPI-based workflow engine design			Real time performance data capture			metadata capture			data integrity assurance				

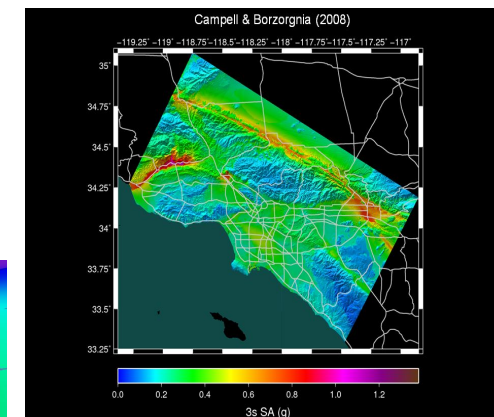
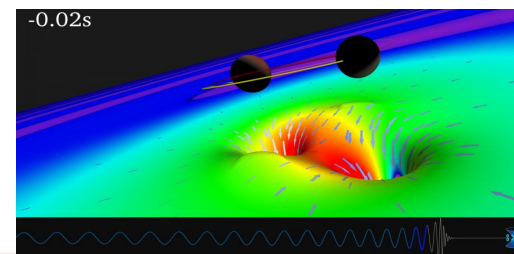
Pegasus
in practice

- ▶ Laser Interferometer Gravitational Wave Observatory (LIGO) develops large scale analysis pipelines used for gravitational wave detection.
- ▶ Southern California Earthquake Center (SCEC) CyberShake project generates hazard maps using hierarchical workflows .
- ▶ The XENONnT project uses Pegasus for processing and monte carlo workflows, searching for dark matter

The XENONnT
detector



LIGO
observation
of colliding
black holes

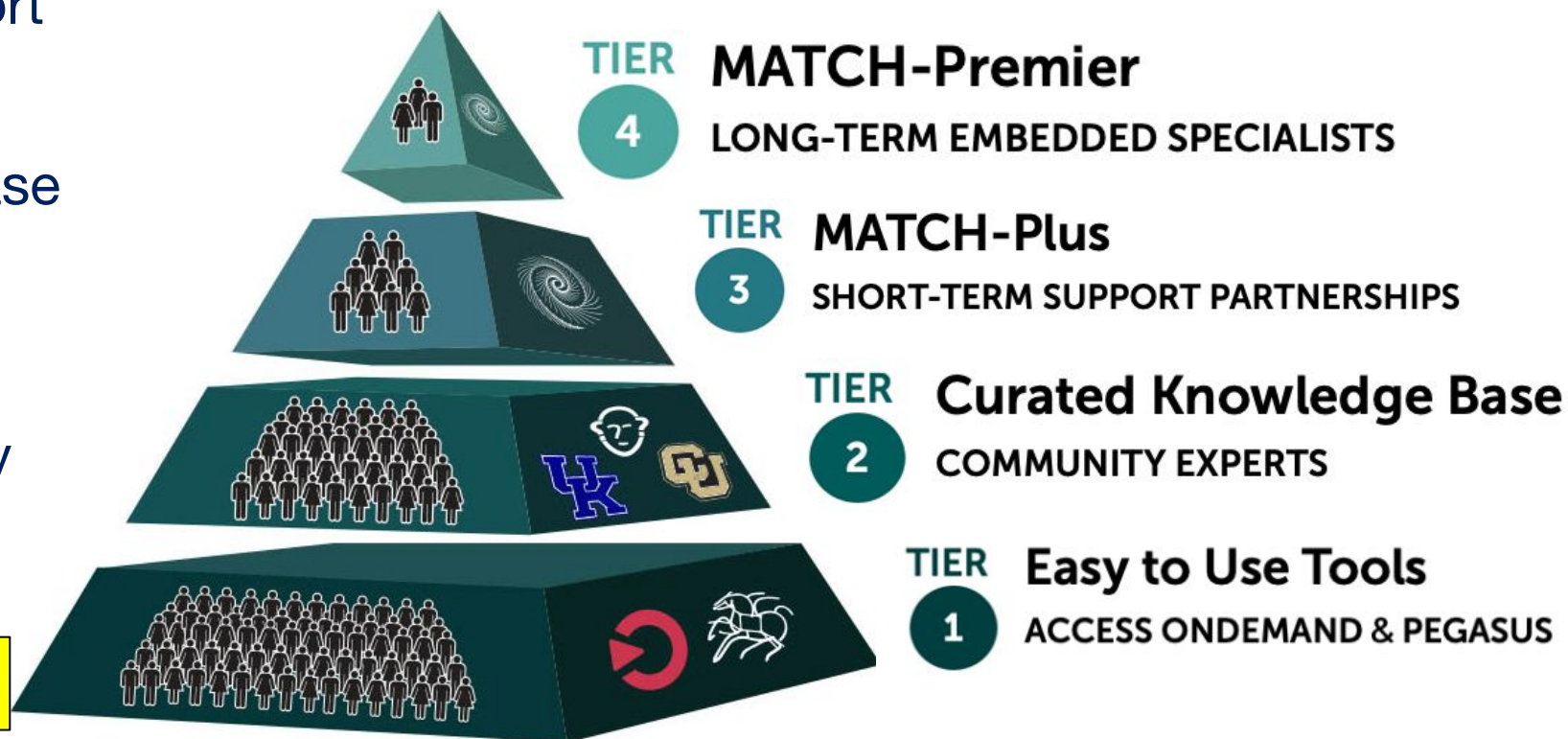


Hazard map indicating maximum amount of shaking at a particular geographic location generated from SCEC's CyberShake Pegasus workflow



ACCESS Research Support Services

- Enable innovative research through equitable and scalable support
- Four tiers of support
- Tools, growing knowledge base
- Match-making with experts
- Student engagement
- Engagement from community
- CSSN incentives



<https://support.access-ci.org>



ACCESS Pegasus

- Open OnDemand
- Jupyter
- Pegasus
- HTCondor AP/CM
- CILogon / ACCESS IdP

<https://pegasus.access-ci.org>



Dashboard - ACCESS Pegasus

<https://pegasus.access-ci.org/pun/sys/dashboard>

ACCESS Pegasus Apps Files Clusters Interactive Apps ACCESS My Interactive Sessions Help Logged in as ryrng Log Out

Jupyter Notebook (create/manage workflows) System Installed App

Local Shell Access System Installed App

Getting Started

Documentation

- ACCESS Pegasus Overview
- Detailed ACCESS Pegasus documentation
- Pegasus User Guide

Quickstart

1. Start an interactive Jupyter session. For the sample workflows, 24 hours runtime is fine. For

My Interactive Sessions - ACCESS Pegasus-Examples/Art... OrcaSound - Jupyter Notebook

<https://pegasus.access-ci.org/node/pegasus.access-ci.org/39992/notebooks/ACCESS-Pegasus-Examp...>

Jupyter OrcaSound Last Checkpoint: Last Friday at 5:33 PM (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run

Workflow

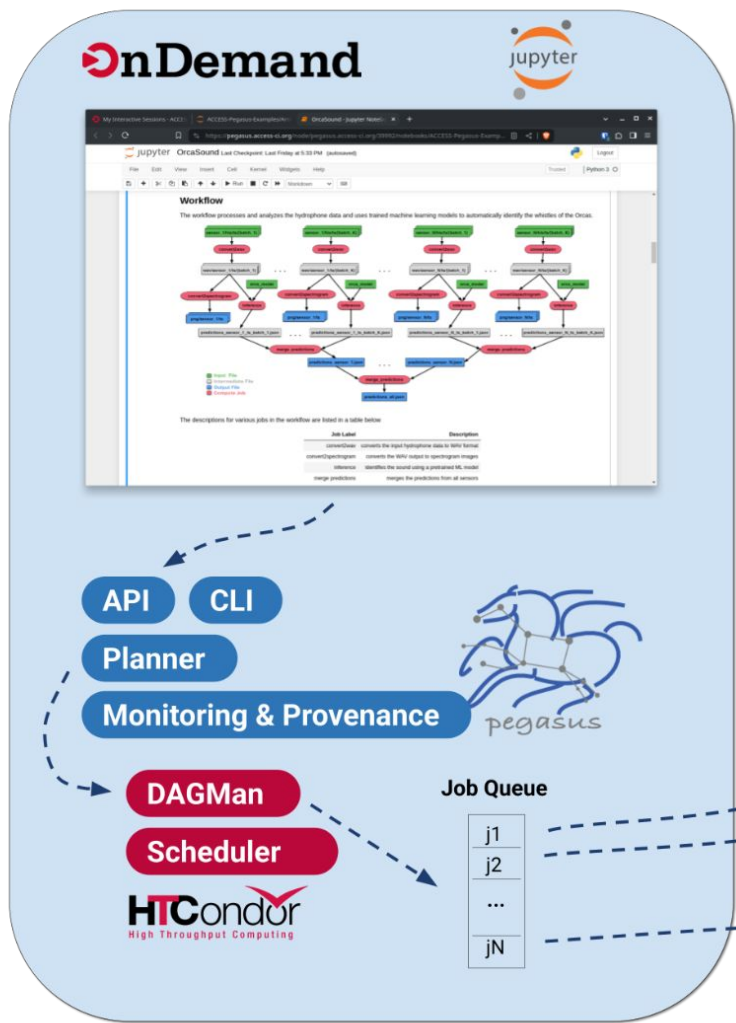
The workflow processes and analyzes the hydrophone data and uses trained machine learning models to automatically identify the whistles of

Legend:

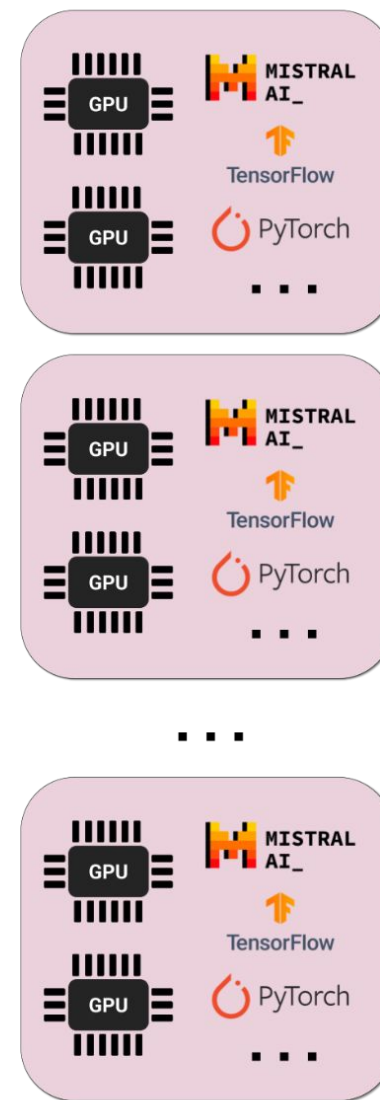
- Input File (Green)
- Intermediate File (Grey)
- Output File (Blue)
- Compute Job (Red)

The descriptions for various jobs in the workflow are listed in a table below

Job Label	Description
convert2wav	converts the input hydrophone data to WAV format
convert2spectrogram	converts the WAV output to spectrogram images
inference	identifies the sound using a pretrained ML model
merge_predictions	merges the predictions from all sensors



pegasus.access-ci.org



TestPool

Cloud

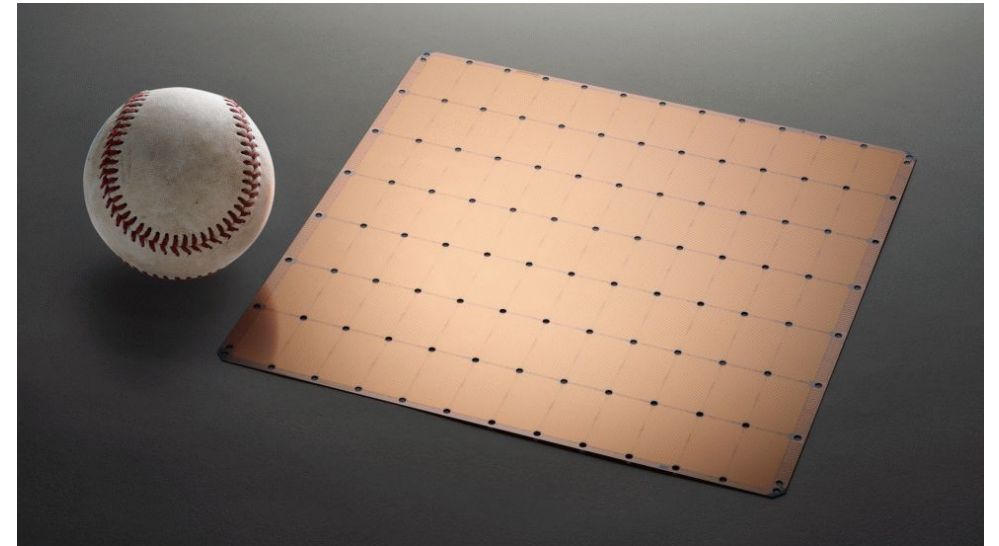
HPC



AI Workflows on PSC Neocortex

Enable complex workflows with mixed job types: ML jobs to Neocortex, and non-ML jobs to Bridges2

- Powerful AI computational resource featuring 2 Cerebras CS-2 systems
- Developed an exemplar Cerebras modelzoo training workflow using TensorFlow



Purdue Anvil Composable Subsystem

- Kubernetes / Rancher
- CILogon (allocation required)
- JupyterLab configured with Slurm, shared filesystems
- HTCondor, batch GAPH, Pegasus



Welcome to the Anvil Notebook Service

Login



The image features decorative geometric patterns in the corners. The top-right and bottom-left corners contain clusters of shapes including circles, triangles, and semi-circles in shades of teal, yellow, and orange. Some shapes are filled with concentric lines. The bottom-right corner has a faint, light blue geometric pattern.

EPs: Resource Provisioning

Resources: Multi-core Compute

Anvil (Purdue) — 1,000 AMD Milan nodes, 128 cores per node, large memory nodes available

Bridges-2 (PSC) — 504 AMD Rome nodes, 128 cores per node, large memory nodes available; extreme memory (4 TB) nodes allocated separately

DARWIN (U Delaware) — Analysis-oriented AMD Rome nodes with 0.5 TB, 1 TB, and 2 TB memory options

Delta (NCSA) — 124 AMD Milan nodes, 128 cores per node

Expanse (SDSC) — 728 AMD Rome nodes, 128 cores & 1 TB NVMe per node

KyRIC (U Kentucky) — Five large-memory (3 TB, 6 TB) nodes, 300 TB storage

Stampede 3 (TACC) — 1,858 Intel Xeon CPU Max nodes

Resources: GPU Computing

Anvil GPU (Purdue) — 16 nodes, 4 NVIDIA A100 GPUs each

Bridges-2 GPU (PSC) — 24 nodes, 8 NVIDIA V100 GPUs & 7.68 TB NVMe per node

Neocortex (PSC) — 2 Cerebras CS-2 servers each with a Cerebras Wafer Scale Engine (WSE-2)

DARWIN GPU (U Delaware) — Large-memory nodes with three different GPU architectures: AMD MI50, NVIDIA T4 & V100

Delta GPU (NCSA) — 4 node configs: 100 nodes w/ 4x A100s; 100 w/ 4x A40 GPUs; five w/ 8x A100s; one w/ 8x AMD MI100 GPUs

Expanse GPU (SDSC) — 52 nodes, 4 NVIDIA V100 GPUs each

Jetstream2 (Indiana U) — 90 nodes with 4x A100 GPUs

Resources: Novel Computing

ACES (Texas A&M U) — Composable PCIe fabric with Intel Sapphire Rapids cores, Graphcore IPU, NEC Vector Engines, Intel Max GPUs, Intel FPGAs, Next Silicon co-processors, NVIDIA H100 GPUs, Intel Optane memory

FASTER (Texas A&M U) — 180 nodes on a composable fabric, 2x Intel Ice Lake processors each, 260 NVIDIA GPUs (five different architectures)

Jetstream2 (Indiana U) — Cloud environment with AMD Milan nodes, and 90 nodes with 4x A100 GPUs

Hive (Georgia Tech) — 484 Intel Cascade Lake nodes. No allocation necessary for Hive Gateway access!

Ookami (Stony Brook U) — 176 nodes with Riken/Fujitsu A64FX processors; additional nodes with AMD Milan, Thunder X2, and Skylake/V100 architectures

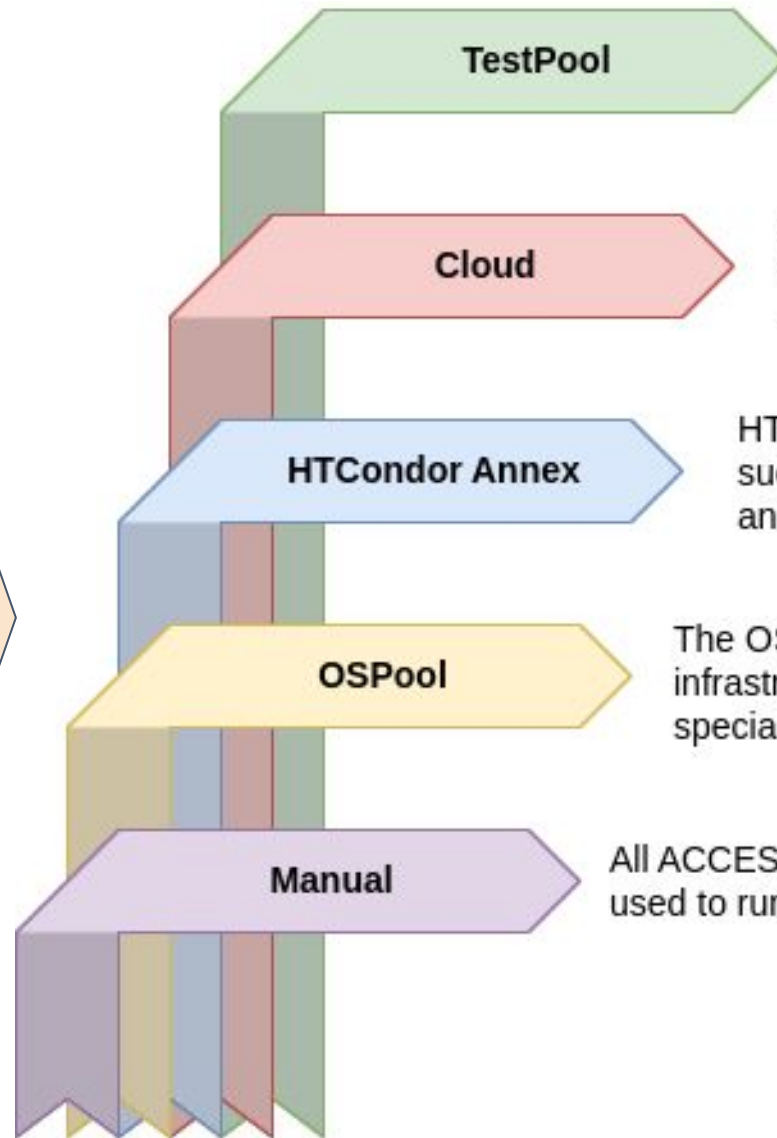
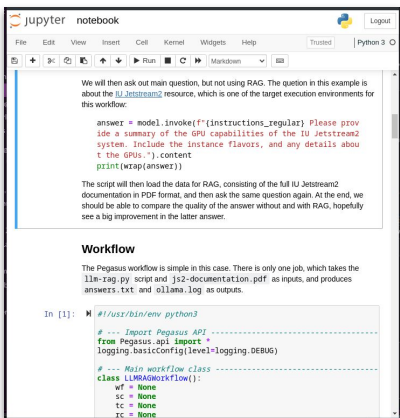
Open Science Pool (PATH) — High-throughput computing environment leveraging fair-share access to contributed compute capacity

Voyager (SDSC) — Habana Gaudi training and inference processors. Kubernetes

Mostly allocation

More diverse catalog then in the past





TestPool

The TestPool is a small resource which is always available. This can be used for development, debugging and running small workflows.

Cloud

For cloud resources, currently IU Jetstream2, a custom image is provided. The image is started with a provided security token, which makes it available to your jobs.

HTCondor Annex

HTCondor Annex enables you to provision compute resources on clusters such as NCSA Delta, SDSC Expanse, PSC Bridges2, Purdue Anvil and PATH Facility

OSPool

The OSG Open Science Pool (OSPool) will run jobs in a distributed infrastructure. You do not need to provision these resources, but there are special rules and capabilities.

Manual

All ACCESS Pegasus users are automatically issued a token which can be used to run glideins on any resource.

Compute Jobs



TestPool

Helps users with an ACCESS account but no allocation explore the capability

Small amount of compute resources attached to pegasus.access-ci.org

- CPU: 32 cores, 128 GB RAM, 256 GB disk
- GPU: 32 cores, 2 GPUs, 128 GB RAM, 256 GB disk
- Hosted on IU Jetstream2, auto-provisioned when needed

Always available, **no allocation needed**

Can be used for quick turnaround jobs

- workflow development and debugging
- tutorials (not all users might have an allocation at the time of the tutorial)
- classroom



Cloud

IU JetStream2

Provided VM image

Users have to add pegasus.access-ci.org username and token in the cloud-init yaml

Instances self-terminates when there are no more jobs



Boot Script

This **cloud-init** ⓘ config describes how to provision the instance. It's provided here to permit specific changes in rare circumstances; please modify it cautiously.

⚠ By editing this it's possible to break various Exosphere features like web desktop, web shell, usage graphs, setup status, etc.

```
#cloud-config
users:
  - default
  - name: exouser
    shell: /bin/bash
    groups: sudo, admin
    sudo: ['ALL=(ALL) NOPASSWD:ALL'] {ssh-authorized-
keys}
ssh_pwauth: true
package_update: true
package_upgrade: {install-os-updates}
packages:
  - git(write-files)
bootcmd:
  - /opt/ACCESS-Pegasus-Jetstream2/bin/vm-conf alice
  aabbcc...
runcmd:
  - echo on > /proc/sys/kernel/printk_devkmsg || true
# Disable console rate limiting for distros that use
kmsg
  - sleep 1 # Ensures that console log output from
any previous command completes before the following
command begins
```



HTCondor Annex



- Bring your own HPC allocation
- Semi-managed, submits glideins via SSH.
 - A glidein can run multiple user jobs - it stays active until no more user jobs are available or until end of life has been reached, whichever comes first.
 - A glidein is partitionable - job slots will dynamically be created based on the resource requirements in the user jobs. This means you can fit multiple user jobs on a compute node at the same time.
 - A glidein will only run jobs for the user who started it.
- Documentation: <https://htcondor.org/experimental/ospool/byoc/>

```
$ htcondor annex create --nodes 1 --lifetime 7200 \  
--project sta230005p --gpu-type v100-16 $USER GPU@bridges2
```

delta

cpu
cpu-interactive
gpuA100x4
gpuA100x4-preempt
gpuA100x8
gpuA40x4
gpuA40x4-preempt

stampede2

normal
development
skx-normal

expanse

compute
gpu
shared
gpu-shared

anvil

wholenode
wide
shared
gpu
gpu-debug

bridges2

RM
RM-512
RM-shared
EM
GPU
GPU-shared

path-facility

cpu



pegasus-glidein

- Simple glidein which can be run anywhere, as long as you have outbound network connectivity
- Ties in well with ACCESS Pegasus
- <https://github.com/pegasus-isi/pegasus-glidein>

```
#!/bin/bash
#SBATCH --job-name=glidein
#SBATCH --nodes=10
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=48
#SBATCH --time=24:00:00

curl -o pegasus-glidein https://raw.githubusercontent.com/pegasus-isi/pegasus-glidein/main/pegasus-glidein
chmod a+x pegasus-glidein
srun ./pegasus-glidein -c pegasus.access-ci.org \
                        -t mytoken \
                        -s 'Owner == "myusername"'
```



Outreach and Workshops



AI Unlocked: Empowering Higher Ed through Research and Discovery



Getting Started Guide for AI Institutes: SAIL 2023

Cyberinfrastructure (CI) resources and support services for your research needs.

Join the Affinity Group

The AI/CI affinity group is a gathering place for AI researchers to find curated information about using ACCESS resources for AI applications and research.

<https://support.access-ci.org/affinity-group/s/ai-institutes-cyberinfrastructure>

Request an Allocation

Most allocation requests are approved within one business day. Get started with exploring the various resources and upgrade your allocation when your resource needs start to intensify.

<https://allocations.access-ci.org>

Get Support

Support is provided both in a self-service format (Knowledge Base, Ask CI), and as concierge-level services that pair your project with experts.

<https://support.access-ci.org>

Recommended Resources

Anvil GPU (Purdue) — 16 nodes, 4 NVIDIA A100 GPUs each

Bridges-2 GPU (PSC) — 33 nodes, 8 NVIDIA V100 GPUs & 7.68 TB NVMe per node

Delta GPU (NCSA) — 4 node configs: 100 nodes with 4x A100s; 100 with 4x A40 GPUs; five with 8x A100s; one with 8x AMD MI100 GPUs

DeltaAI (NCSA) — *Coming Soon*

Expanse GPU (SDSC) — 52 nodes, 4 NVIDIA V100 GPUs each

DARWIN GPU (U Delaware) — Large-memory nodes with three different GPU architectures: AMD MI50, NVIDIA T4, & V100

Browse all available resources at:
allocations.access-ci.org/resources

Have questions? Get in contact with ACCESS staff:
<https://support.access-ci.org/open-a-ticket>

- NAIRR: AI Unlocked
 - Regional versions coming soon: RMACC, Kentucky, Ohio, SoCal
- Duke IEEE x ACCESS
- PEARC
- USRSE
- 5 tutorials across time zones

