Introducing the PATh Facility

OSG All-Hands Meeting,

March 2022
Brian Bockelman

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

- Starting in 2020, the **Partnership to Advance Throughput Computing** (PATh) brought together
  - The OSG Consortium
  - The Center for High Throughput Computing (CHTC)

with a goal providing

> services and technologies for distributed high-throughput computing (dHTC)!

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

OSG

HT CENTER FOR HIGH THROUGHPUT COMPUTING

Technologies and Services

Technologies

- HTCondor Software Suite
- Access Points
- Backfill containers
- OSDF containers

Services

Technologies and Services

Technologies

- HTCondor Software Suite
- Access Points
- Backfill containers
- OSDF containers

Services

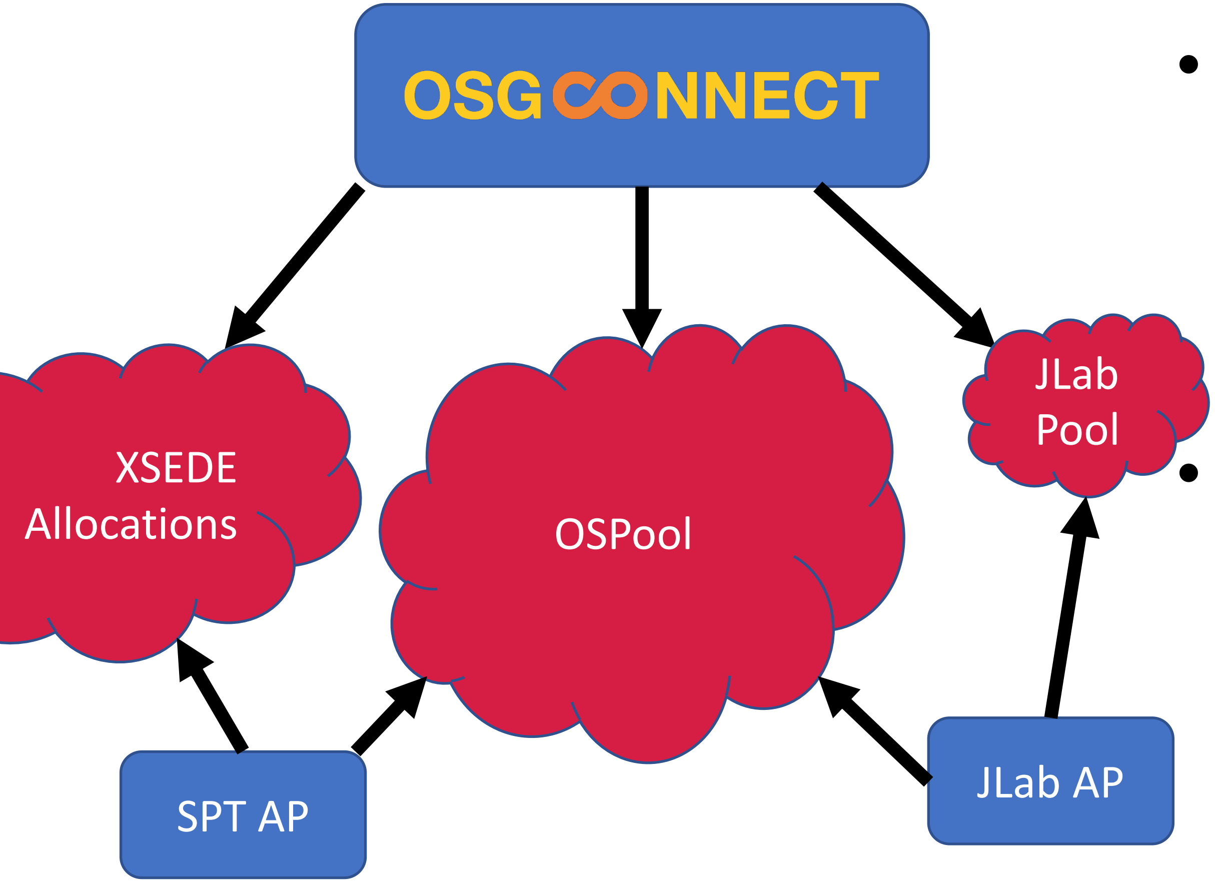
- OSPool
- OSG Connect
- OSDF
- Resource Provisioning
- Hosted CEs
- Resource Pools
- Many more! Rucio, FTS, CVMFS, Monitoring, Accounting, Collaboration data services.

OSG CONNECT

OSG Connect is an instance of an Access Point service, operated by OSG.

- OSG Connect at UChicago is an example Access Point (AP).
  - Users can place their workloads and data at OSG Connect,
  - Then OSG Connect can attach to different resource pools to execute the workloads.
- The AP serves as a portal to different resource pools.
- At the heart is the HTCondor submit service. But also:
  - Data movement and management.
  - User/group management.
  - Unix account management.
  - Integration with resource provisioning.

The Open Science Pool (OSPool)



- The OSPool is a pool of resources operated by the OSG.
  - An AP – like OSG Connect –can attach to the OSPool and utilize its resources.
- In general, the resource pools (OSPool) and APs (OSG Connect) are independent, top-level entities.
  - For APs, this forms the basis of the idea "submit locally, run globally".

The Open Science Pool (OSPool)

FREE COMPUTERS*!

FREE COMPUTERS*!

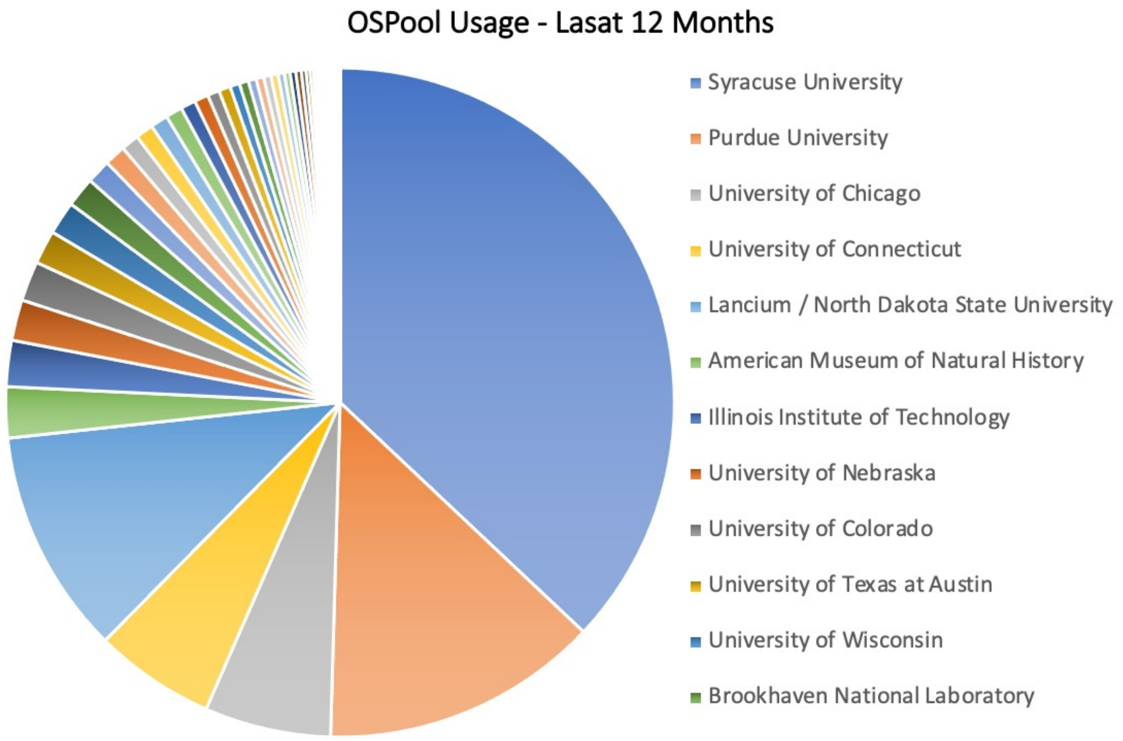
FREE COMPUTERS*!

FREE COMPUTERS*!

FREE COMPUTERS*!

FREE COMPUTERS*!

FREE COMPUTERS*!



*OSPool resources are donated from many resources and subject to their rules and availability. Computers range between 12 days and 12 years old.

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

OSG

HT CENTER FOR HIGH THROUGHPUT COMPUTING

Managing Computing Resources

- OSPool operates by fairshare.
  - Resources are managed by the OSG Executive Director ->
  - Resources are entrusted (opportunistic or donated) to OSG with the understanding OSG will do the best science with them as possible.
    - However, there's a limit to the quality of service we can offer.
- There are other mechanisms to manage resources:
  - XSEDE provides allocations to specific machines.
  - AWS provides a "pay $$$ as you go" model.
  - CloudBank manages cloud compute credits that are part of your NSF award.

CloudBank

Credit Accounts for HTC

- Motivation: What does a ‘credit account service’ look like for an HTC system?
- Idea:
  - Researchers can receive a a credit account with a certain amount of “funny money” in the account.
    - Different currencies may be available for non-fungible resource types (GPU vs CPU).
  - Researchers can spend this on a range of HTC resources – not tied to a single site or resource type.
    - We compute the “charge” based on types of resources used and
  - Have built-in functionality in HTCSS to support this service.
- Observation: To really test this idea, we need production-quality hardware that is interesting to researchers.

Technologies, Services, and Resources

Technologies

- HTCondor Software Suite
- Access Points
- Backfill containers

Credit Accounts in HTCSS

Services

- OSPool
- OSG Connect
- OSDF
- Resource Provisioning

Credit Account Service

Resources

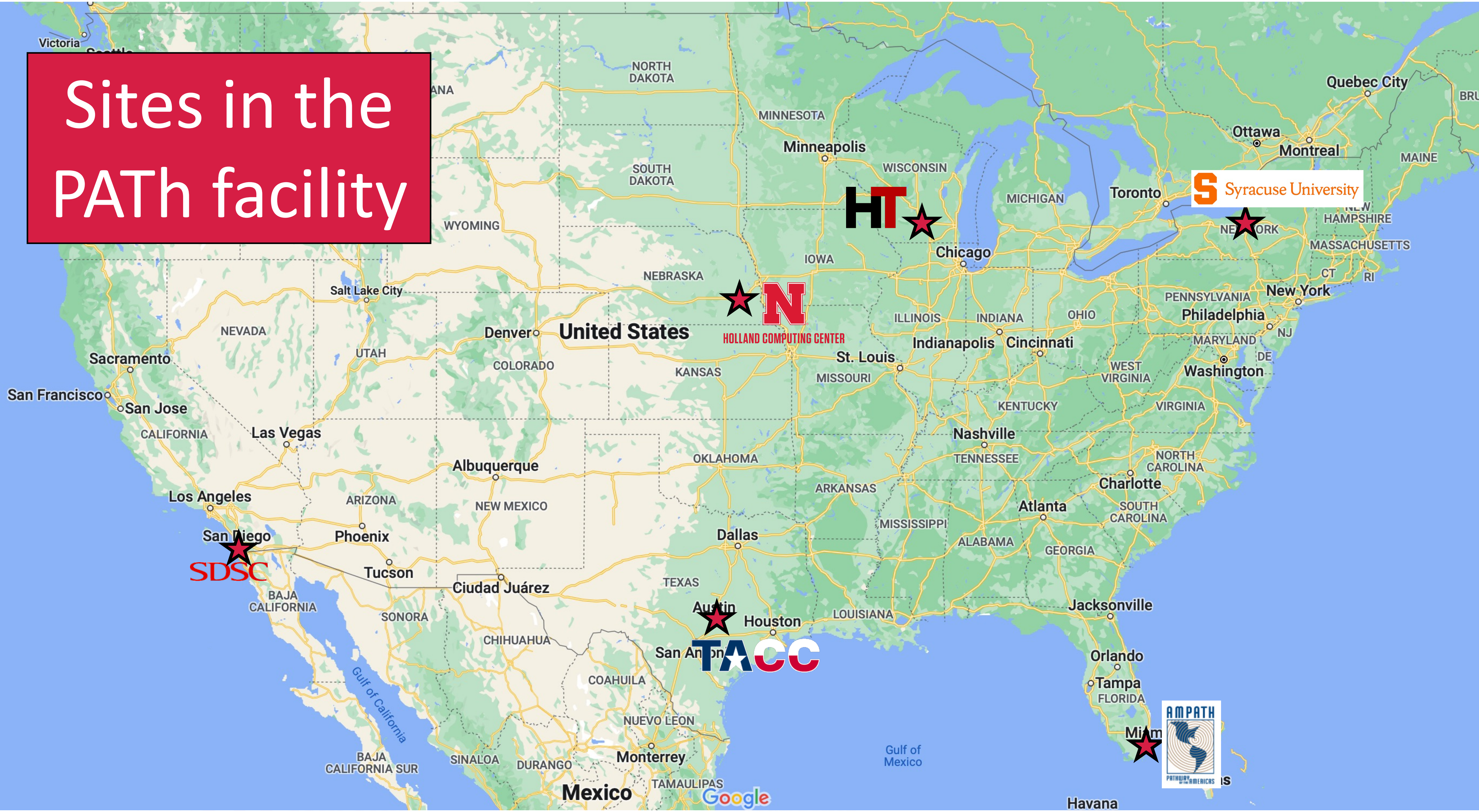
- PATh Facility

A Distributed Facility for dHTC

https://path-cc.io/facility/

- To complement the credit accounts for dHTC, we put together a uniquely dHTC resource.
  - No shared filesystem, physically distributed, reflects what can be found in the OSPool.
- The PATh Facility is a distributed resources across six sites, meant for dHTC workflows in support of open science.
  - Four sites (Florida, Syracuse, Nebraska, and Wisconsin) will get hardware funded through the PATh project itself.
  - Two sites are funded as part of extensions to existing resources (Expanse at SDSC, Stampede2 at TACC).
- The Facility is uniquely distributed. For example, hardware is located "on the (network) path" at AMPATH in Miami.
- Unlike on the OSPool, "we make the rules". Longer jobs are more practical as we control whether the remote site does preemption.

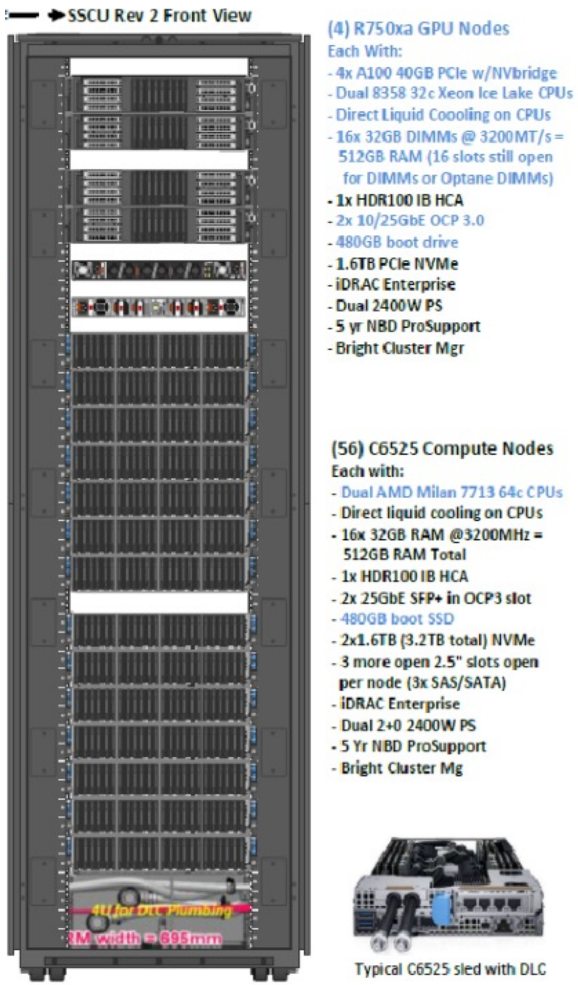
Sites in the PATh facility



Quick Hardware Overview

- The PATh-owned hardware consists of:
  - CPU: 64 AMD EPYC 7513 cores (2.6GHz), 256GB RAM, 1.6TB NVMe.
    - 32 nodes @ Nebraska, 32 @ Syracuse, 4 @ AMPATH, 4 @ Wisconsin (primarily for debug / development).
  - GPU: 4 x A100 GPU servers, 512GB RAM, 1.6TB NVMe.
    - One at Nebraska, Syracuse, and AMPATH.
  - Service: 60TB NVMe; shared filesystem to other local hosts.
  - All connected at 100GbE.
- TACC resources consist of Stampede2, which was recently extended with 224 Intel Platinum 8380 “Ice Lake” serves.

The SDSC resources are two racks of this



(4) R750xa GPU Nodes
Each With:
- 4x A100 40GB PCIe w/NVbridge
- Dual 8358 32c Xeon Ice Lake CPUs
- Direct Liquid Coooling on CPUs
- 16x 32GB DIMMs @ 3200MT/s = 512GB RAM (16 slots still open for DIMMs or Optane DIMMs)
- 1x HDR100 IB HCA
- 2x 10/25GbE OCP 3.0
- 480GB boot drive
- 1.6TB PCIe NVMe
- iDRAC Enterprise
- Dual 2400W PS
- 5 yr NBD ProSupport
- Bright Cluster Mgr

(56) C6525 Compute Nodes
Each with:
- Dual AMD Milan 7713 64c CPUs
- Direct liquid cooling on CPUs
- 16x 32GB RAM @3200MHz = 512GB RAM Total
- 1x HDR100 IB HCA
- 2x 25GbE SFP+ in OCP3 slot
- 480GB boot SSD
- 2x1.6TB (3.2TB total) NVMe
- 3 more open 2.5" slots open per node (3x SAS/SATA)
- iDRAC Enterprise
- Dual 2+0 2400W PS
- 5 Yr NBD ProSupport
- Bright Cluster Mg

Typical C6525 sled with DLC

Between the PATh, Expanse, and Stampede2 projects,

Researchers can get credits accessing

> 35,000 modern AMD / Intel cores

44 A100 GPUs

At 6 sites, including one in a R&E network colo

All as part of the PATh Facility!

Technology “Under the Hood”

- Each of the 6 sites will operate as an independent Kubernetes cluster.
  - PATh will run 4 of these clusters; we’ll be a tenant in the other 2.
- PATh will run the central manager and accounting services on the Tiger & River Kubernetes clusters.
  - Initially, the AP will be run at Wisconsin to allow rapid development/changes on the accounting services.
- All service configurations are kept in a single git repository and deployed by flux; a single `git push` can deploy across the facility.
- Per-pod storage is allocated by OpenEBS’s LVM operator for local storage and Mayastor for shared/site-level storage.
- Beyond worker node pods, we eventually plan for OSDF caches at each site.
- There will be some unique network challenges. For example, the AMPATH site will start with only IPv6 connectivity.

PATh Credit Accounts

- Unlike the OSPool, the OSG Consortium doesn’t make decisions about PATh facility resource allocations.
    - Rather, NSF creates and hands out credits.
    - Do not worry: if there are idle resources, we can push them into the OSPool!
- Two mechanisms (so far):
    - The 2021 CSSI solicitation from OAC allowed PIs to request credits as part of their proposal.
    - A 2022 DCL provides another mechanism for a broad range of NSF programs.
- With the upcoming projects in the ACCESS program, we believe NSF will continue to innovate with how credits are assigned to researchers.

National Cyberinfrastructure Coordination Services

- I do not see the PATh facility as an independent entity.
  - Rather, it's an example of a service fitting in NSF's National Cyberinfrastructure Coordination Services.
- It's not meant to be the largest or the fastest – rather it's meant to help bootstrap the credit-based accounting approach within PATh.
  - And serve as a vanguard, exploring different resource management mechanisms.

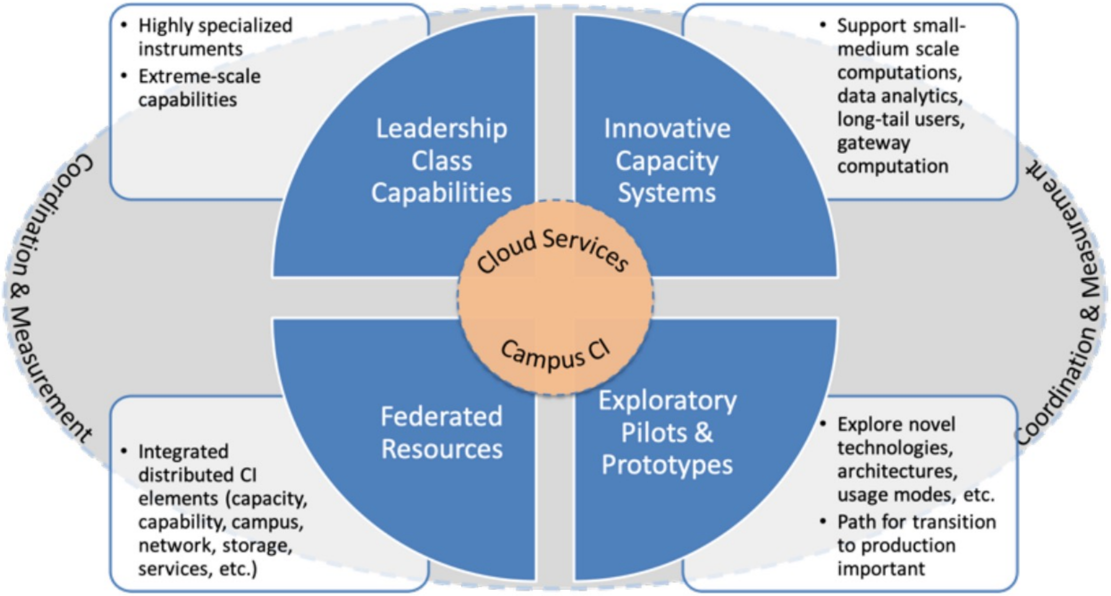

• Highly specialized instruments
• Extreme-scale capabilities
Leadership Class Capabilities
Innovative Capacity Systems
• Support small-medium scale computations, data analytics, long-tail users, gateway computation
Cloud Services
Campus CI
Federated Resources
Exploratory Pilots & Prototypes
• Integrated distributed CI elements (capacity, capability, campus, network, storage, services, etc.)
• Explore novel technologies, architectures, usage modes, etc.
• Path for transition to production important
Coordination & Measurement
Coordination & Measurement

Figure 5: Elements of a balanced computational ecosystem.

Figure reproduced from OAC's "Transforming Science through Cyberinfrastructure"

PATh Facility – Looking Forward

- The software and services for credit accounts have initial prototypes deployed.
- The PATh facility is rapidly becoming available in the first half of 2022:
  - PATh-owned hardware is online at Wisconsin and available for testing jobs and activities via the OSPool.
  - Hardware will be shipped to remote destinations throughout spring.
  - Expanse/SDSC hardware is expected in early summer.
  - TACC resources are expected to be available in a similar timeframe.
- PIs can request credits now by contacting program officers in the participating programs.

Initial PATh-owned hardware, destined for Syracuse

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

OSG

HT CENTER FOR HIGH THROUGHPUT COMPUTING

Acknowledgements

MORGRIDGE
INSTITUTE FOR RESEARCH
RESEARCH COMPUTING

This project is supported by National Science Foundation under Cooperative Agreement OAC-2030508. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

NSF

For more information, visit https://path-cc.io/facility/

PATh PARTNERSHIP to ADVANCE THROUGHPUT COMPUTING

HT CENTER FOR HIGH THROUGHPUT COMPUTING

OSG