High Throughput Data-Intensive Computing in the Era of Ever Larger Supercomputers

Miron Livny

John P. Morgridge Professor of Computer Science
Wisconsin Institutes for Discovery
University of Wisconsin-Madison

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

In **1996** I introduced the distinction between High **Performance** Computing (**HPC**) and High **Throughput** Computing (**HTC**) in a seminar at NASA Goddard Flight Center and a month later at European Laboratory for Particle Physics (**CERN**).

In June of 1997 HPCWire published an interview on High Throughput Computing.

```
HIGH THROUGHPUT COMPUTING: AN INTERVIEW WITH MIRON LIVNY           06.27.97
by Alan Beck, editor in chief                                       HPCwire
============================================================================

  This month, NCSA's (National Center for Supercomputing Applications)
Advanced Computing Group (ACG) will begin testing Condor, a software system
developed at the University of Wisconsin that promises to expand computing
capabilities through efficient capture of cycles on idle machines. The
software, operating within an HTC (High Throughput Computing) rather than a
traditional HPC (High Performance Computing) paradigm, organizes machines
```

From the
I-W
Na
Te

C

environments can provide them per second. We call this high-throughput computing, or HTC, in contrast with classic high-performance computing, or HPC. HTC is being pioneered by the University of Wisconsin's Condor project, which organizes desktop workstations from a buildingwide to a campuswide (or beyond) computing environments with thousands of processors, creating a high-throughput facility [8].

The Condor resource management system employs a novel approach to HTC based on a layered architecture and preemption/resume scheduling.

The Grid coulc
into a national
metacomputer
everyone, but f
comes integrati
networking, di
computing, sch
Web and Java
programming, security.

BY ASSEMBLING NATIONWIDE TEAMS OF COMPUTER

Rick Stevens, Paul Woodward, Tom DeFanti, and Charlie Catlett

Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020

"… many fields today rely on **high-throughput computing** for **discovery**."

"Many fields **increasingly** rely **on high-throughput computing**"

AUTHORS

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020

"Recommendation 2.2. NSF should (a) … and (b) broaden the accessibility and utility of these large-scale platforms by allocating **high-throughput** as well as high-performance workflows to them."

AUTHORS

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

Towards a Leadership-Class Computing Facility - Phase 1

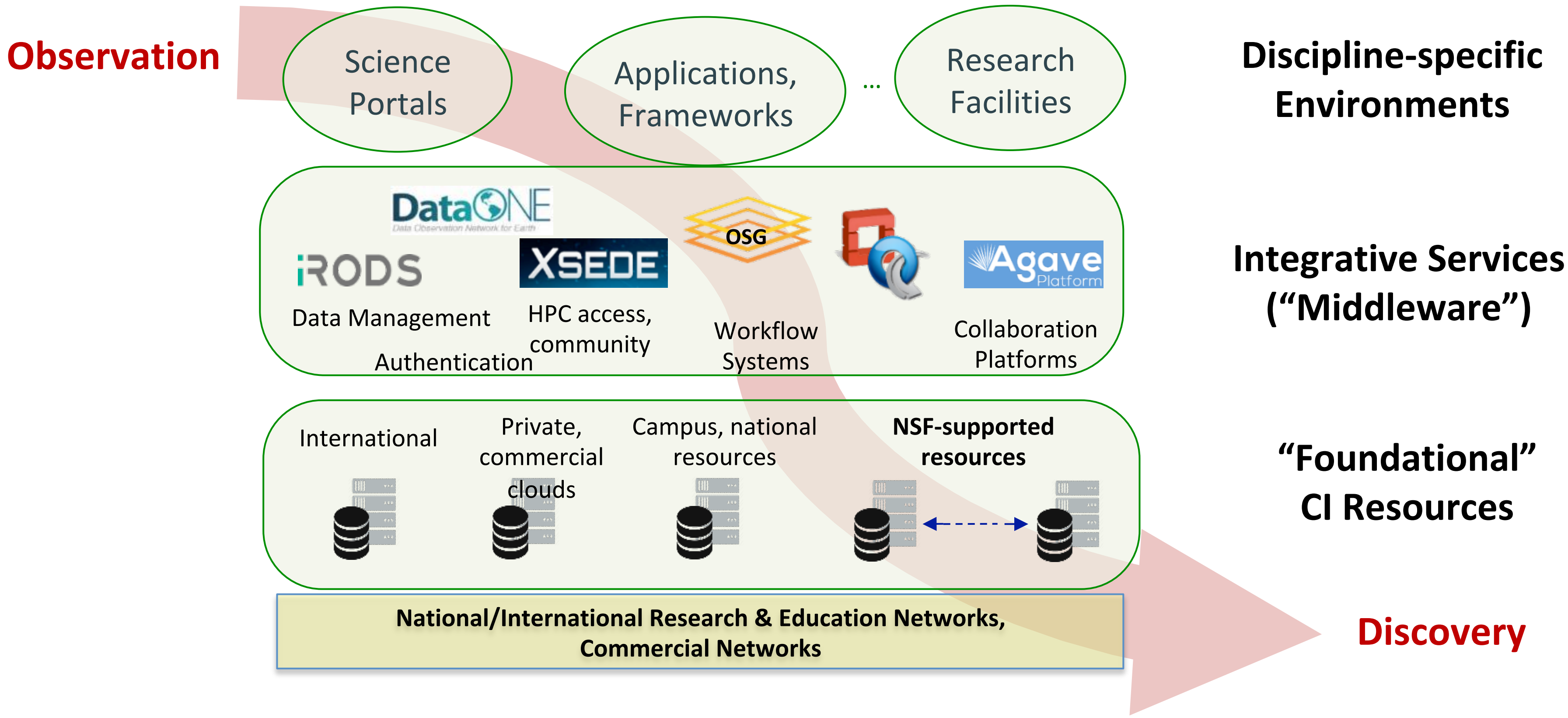
PROGRAM SOLICITATION
NSF 17-558

National Science Foundation

Directorate for Computer & Information Science & Engineering
Division of Advanced Cyberinfrastructure

The submitted performance analysis must include a broad range of applications and workflows requiring the highest capabilities in terms of scale (massive number of processors used in a single tightly-coupled application), **high throughput** (massive number of processors used by ensembles), and data analytics (large scale-out workloads for massive data analysis). Performance analysis must include **projected time-to-solution** improvement for all applications running on the proposed Phase 1 system over the existing BlueWaters system.

Dynamic discovery pathways at scale: Architecture view



Overarching Goal

Elasticity (cloud computing) aims at matching the amount of resource allocated to a service (/application/workflow) with the amount of resource it actually requires, avoiding over- or under-provisioning.

[en.wikipedia.org/wiki/Elasticity_(cloud_computing)]

Commercial cloud assume unlimited resources with a well defined cost model while Research Computing operates with bounded resources and a convoluted cost model

Funds to purchase + Funds to lease + Allocations + Fair Share

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

Elasticity @ UW-Madison campus

- The Center for High Throughput Computing (**CHTC**) has been serving the campus for more than 10 years
- Delivered more than **395M** core hours in the past 12 month to researchers with **HTC** workloads from more than **50** departments
- ~10% of these core hours come from off-campus resources – mainly the Open Science Grid (**OSG**)

MORGRIDGE
INSTITUTE FOR RESEARCH

CENTER FOR
HIGH THROUGHPUT
COMPUTING

WISCONSIN
INSTITUTE FOR DISCOVERY

WISCONSIN
INSTITUTE FOR DISCOVERY

Featured Science

Mining the Mind: High Throughput Computir and the Future of Brain Research

February 20, 2014

MORGRIDGE
INSTITUTE FOR RESEARCH

ABOUT RESEARCH OUTREACH NEWS

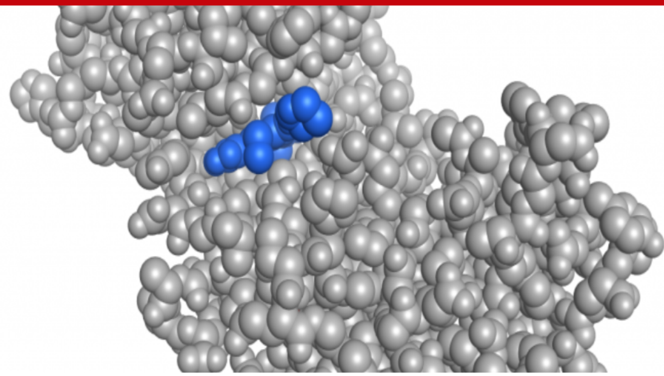
News & Stories > UW botanist harnesses the grid to illuminate crop

UW botanist harnesses the grid to illuminate crop growth

UWMadScience
Behind the science & research that makes the news at UW–Madison

Search

HOME ABOUT NEWS MORE SCIENCE BLOGS



: science » The machines are learning to develop new targeted drug therapies

ıt Posts

ves change color in

The machines are learning to develop new targeted drug therapies

In 24 hours the **CHTC** serves a broad and diverse collection of science disciplines

Fm:	2018-04-18	Total	
To:	2018-04-19	Hours	%Pool
95	Projects	1,316,234	100.0%
1	CMS	379,983	28.9%
2	IceCube	179,903	13.7%
3	Genetics_Pool	77,825	5.9%
4	WID_Biology_Vetsigian	55,985	4.3%
5	Genetics_Payseur	48,931	3.7%
6	BMI_Gitter	48,391	3.7%
7	Biostat	42,618	3.2%
8	Marschfield_Hebbring	33,354	2.5%
9	NutritionalSciences_Parks	26,261	2.0%
10	SmallMolecule_Hoffman	26,226	2.0%
11	Statistics_Ane	23,647	1.8%
12	Physics_Yavuz	22,593	1.7%
13	Chemistry_Berry	21,503	1.6%
14	MIR_Velten	20,793	1.6%
15	BMRB	20,690	1.6%
16	MaterialScience_Morgan	18,860	1.4%
17	EngrPhysics_Sovinec	17,992	1.4%
18	CEE_Wang	15,483	1.2%
19	Chemistry_Schmidt	14,806	1.1%
20	Statistics_YazhenWang	14,499	1.1%

MORGRIDGE
INSTITUTE FOR RESEARCH

CENTER FOR
HIGH THROUGHPUT
COMPUTING

WISCONSIN
INSTITUTE FOR DISCOVERY

Many owners of Resources!

Fm:	2017-03-10	Total	
To:	2017-03-11	Hours	%Pool
79	Projects	1,034,964	100.0%

CAE		CHTC		CS		OSG		WID		SLURM		HEP		COON		DOIT	
Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool
372	0.0%	273,296	26.4%	3,927	0.4%	103,195	10.0%	13,149	1.3%	150,465	14.5%	276,701	26.7%	43	0.0%	11,516	1.1%

SSEC		LMCG		BMRB		WEI		ICECUBE		BIOSTAT		MATH		BIOCHEM		WAISMAN	
Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool	Hours	%Pool
4,771	0.5%	0	0.0%	1,271	0.1%	594	0.1%	143,507	13.9%	43,225	4.2%	0	0.0%	7,987	0.8%	939	0.1%

Resources located in data centers, machine rooms, laboratories, class rooms, desks, …

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

CHTC got from Argonne National Laboratory an allocation of 100K node hours on Cooley.

Cooley Node Configuration

- Architecture: Intel Haswell
- Processors: Two 2.4 GHz Intel Haswell E5-2620 v3 processors per node (6 cores per CPU, 12 cores total)
- GPUs: One NVIDIA Tesla K80 (with two GPUs) per node
- Memory/node: 384GB RAM per node, 24 GB GPU RAM per node (12 GB per GPU)
- FDR Infiniband interconnect
- 345GB local scratch space

How can we make these resources (seamlessly) available to researchers on the UW-Madison campus?

We (Gitter Lab) have a dataset from a UW-Madison biochemistry collaborator where each instance is a mutated form of a protein sequence and a score about what the mutation does to the protein. We want to be able to predict the effects of new combinations of mutations.

The computing side is that we may have roughly 1000 slightly different versions of a deep neural network to test for this problem. Perhaps more than that in the long run, but ~1000 is an approximate batch size. Each of those networks may be able to train on this dataset in 12-24 hours (rough guess) on a Tesla K80 GPU that the Cooley nodes have.

Deploy an HTCondor "Annex" on Cooley with affinity to a specific workflow of Gitter Lab

- Provisioning of Annex is automated
- Annex integrated into the CHTC (HTCondor) environment
- Make sure that workflow completes

These Opportunities (Challenges) are not new!

Claims for "benefits" provided by Distributed Processing Systems

P.H. Enslow, *"What is a Distributed Data Processing System?"* Computer, January 1978

- High Availability and Reliability
- High System Performance
- Ease of Modular and Incremental Growth
- Automatic Load and Resource Sharing
- Good Response to Temporary Overloads
- Easy Expansion in Capacity and/or Function

" … Since the early days of mankind the primary motivation for the establishment of *communities* has been the idea that by being part of an organized group the capabilities of an individual are improved. The great progress in the area of inter-computer communication led to the development of means by which stand-alone processing sub-systems can be integrated into multi-computer *'communities'*. … "

Miron Livny, " *Study of Load Balancing Algorithms for Decentralized Distributed Processing Systems.*",
Ph.D thesis, July 1983.

MORGRIDGE INSTITUTE FOR RESEARCH

HT CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

Networks enable coordination of activities between autonomous entities across trusted connections

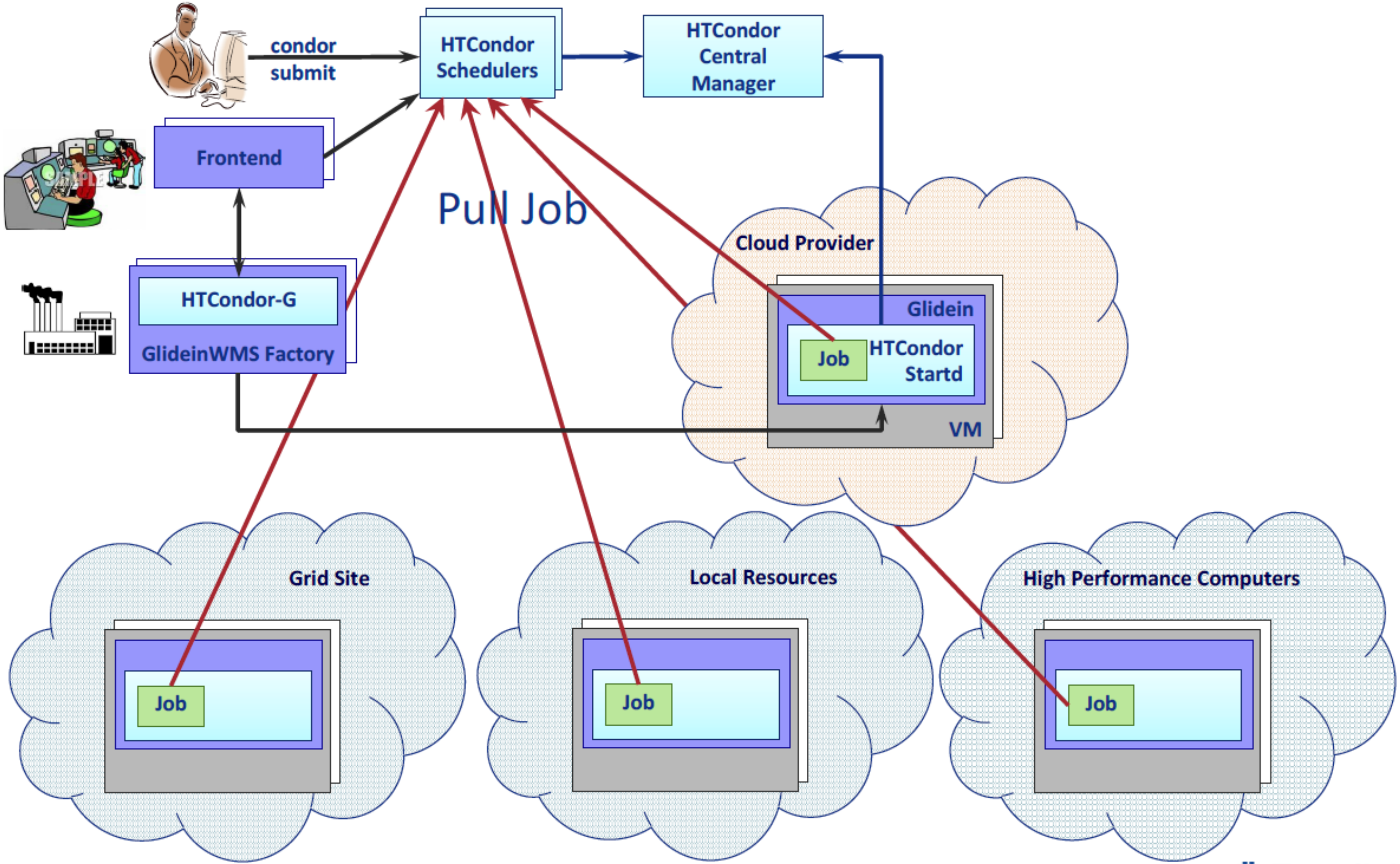
HEPCloud is an R&D project led by the Fermi National Laboratory computing division

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

HEPCloud – glideinWMS and HTCondor



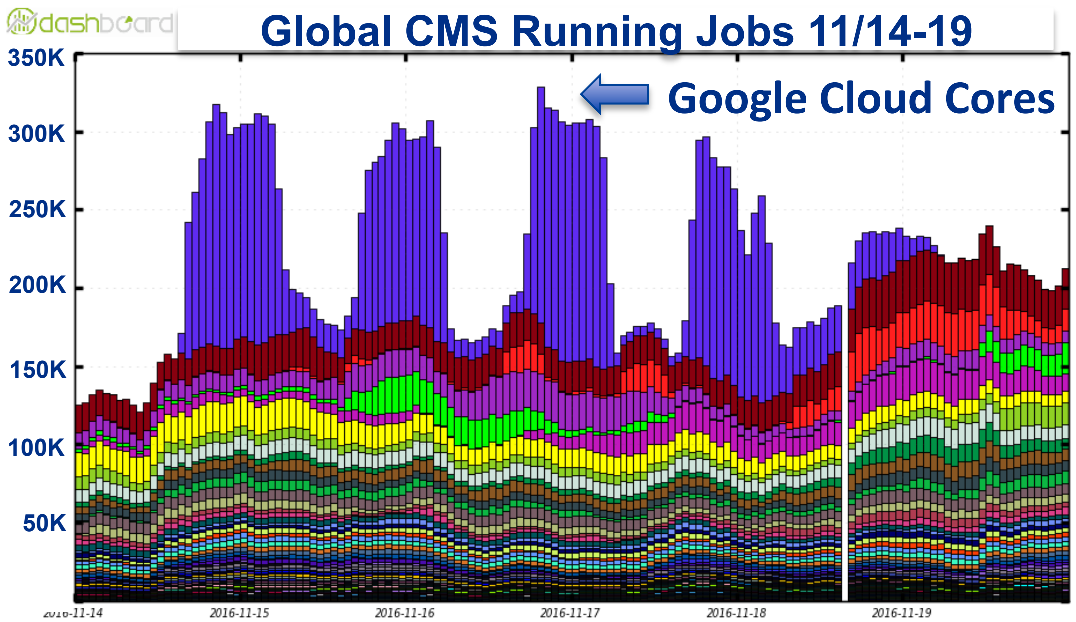
Decision Engine will have to implement **on-the-fly capacity planning** (elasticity) to control acquisition and release of resources

SC16 Demo: On Demand Doubling of CMS Computing Capacity

Joint project

HEPCloud (Fermilab), **HTCondor** (UW-Madison), **Google Cloud**

- **HEPCloud** provisions **Google Cloud** with **HTCondor** in two ways
  - **HTCondor** talks to Google API
  - Resources are joined into **HEP HTCondor** pool
- Demonstrated sustained large scale elasticity (>150K cores) in response to demand and external constraints
  - Ramp-up/down with opening/closing of exhibition floor
  - Tear-down when no jobs are waiting



730,172 jobs consumed 6.35M core hours to produce 205M simulated events (81.8 TB)

Total cost ~$100K

500 TB were placed in Google Cloud in advance. 80TB where moved back to Fermi.

- $8.6k network egress
- $8.5k disk attached to VMs
- $3.5k cloud storage for input data

The Open Science Grid (OSG) national fabric of distribute HTC services

MORGRIDGE
INSTITUTE FOR RESEARCH

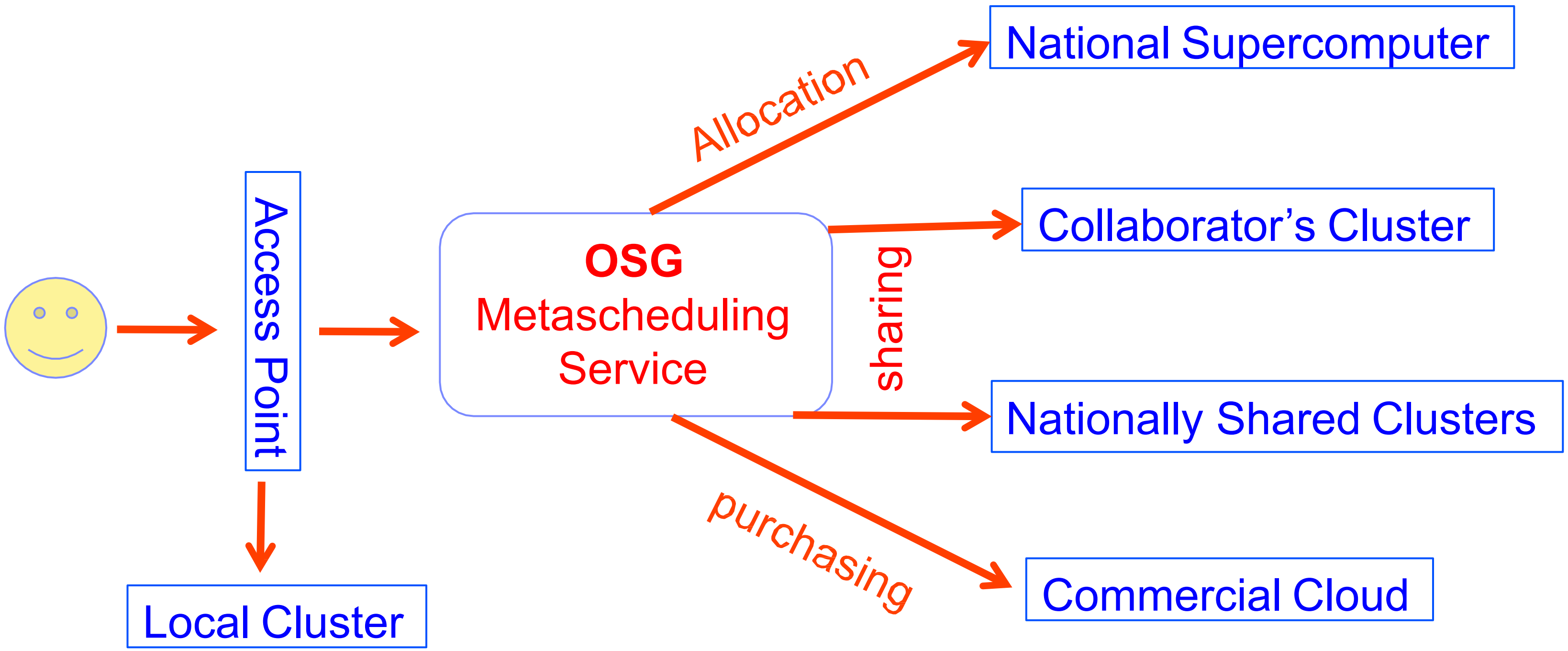
CENTER FOR
HIGH THROUGHPUT
COMPUTING

WISCONSIN
INSTITUTE FOR DISCOVERY

"The members of **OSG** are united by a commitment to promote the adoption and to advance the state of the art of *distributed high throughput computing* (***DHTC***) – *shared utilization of autonomous* resources where all the elements are optimized for maximizing computational throughput."

HTC customers want to **run globally** **(acquire any resource (local or remote) that is capable and for as long as it is willing to run their job/task) while submitting locally (queue and manage their resource acquisition and jobs/tasks (workflows) execution locally)**

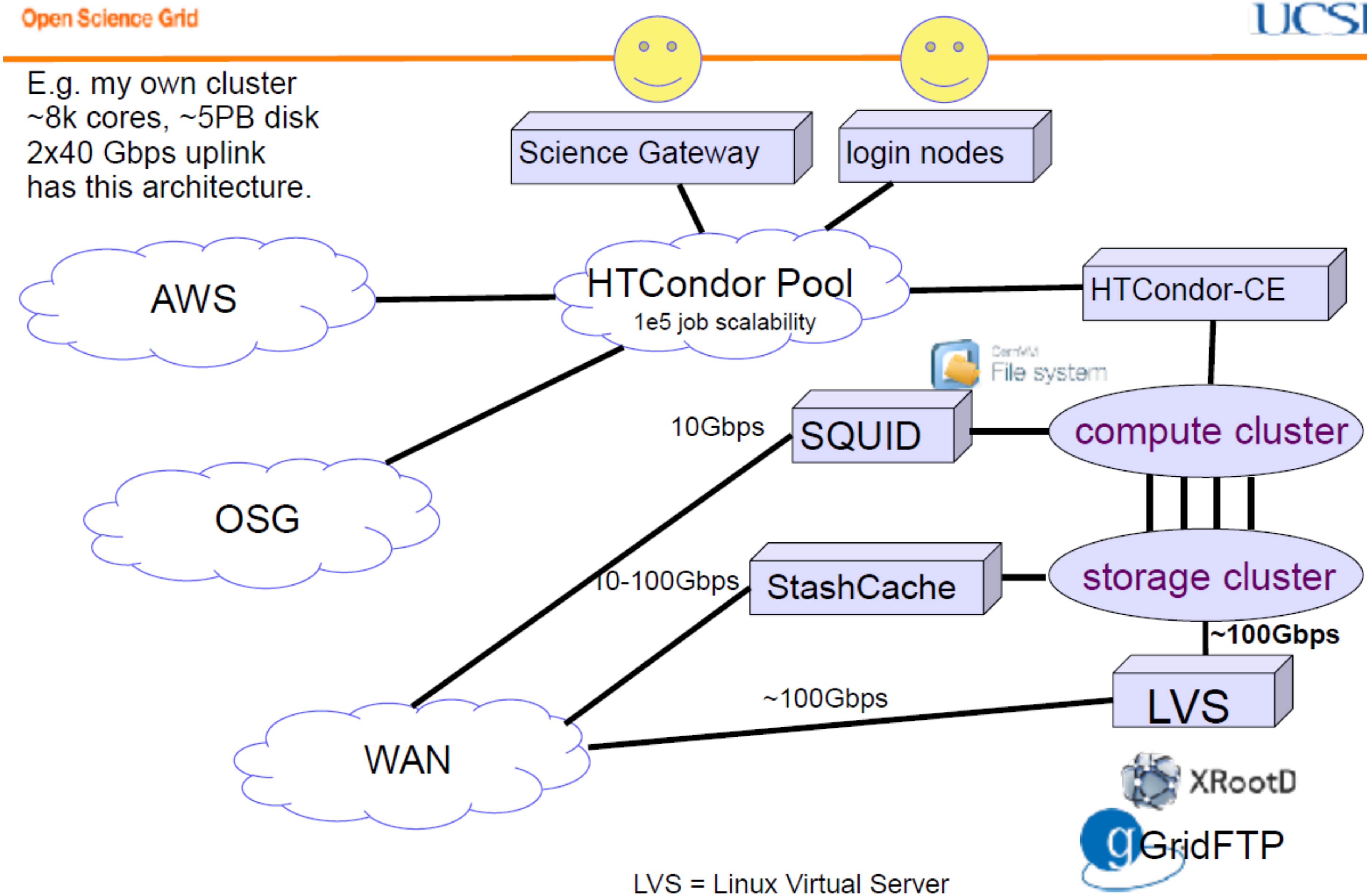
Transparent Computing across different resource types



The Open Science Grid (OSG) integrates computing across different resource types and business models.

An elaborate OSG Site

E.g. my own cluster
~8k cores, ~5PB disk
2x40 Gbps uplink
has this architecture.



LVS = Linux Virtual Server

1.59B core hours in 12 months!

Almost all jobs executed by the **OSG** leverage (HT)**Condor** technologies:

- **Condor-G**
- **HTCondor-CE**
- **Basco**
- **Condor Collectors**
- **HTCondor overlays**
- **HTCondor pools**

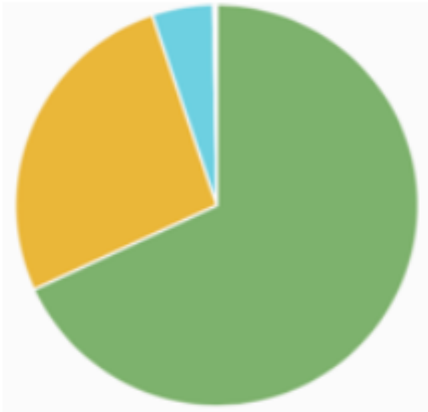
In the last 24 Hours	
336,000	Jobs
4,884,000	CPU Hours
7,553,000	Transfers
677	TB Transfers
In the last 30 Days	
9,229,000	Jobs
138,161,000	CPU Hours
215,046,000	Transfers
23,746	TB Transfers
In the last 12 Months	
135,381,000	Jobs
1,593,317,000	CPU Hours
2,329,401,000	Transfers
218,000	TB Transfers

MORGRIDGE INSTITUTE FOR RESEARCH

HT CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

Core Hours from HPC systems



PSC_Bridges	8748696
Comet	3429688
T3_US_NERSC	630343
Xstream	25918
BlueWaters	7729
Jetstream-CE-1	4226

	values
IBN130001-Plus	8748696
xenon1t	1414289
IceCube	1026836
LIGO	1026436
mu2e	630343

How to integrate Supercomputers into our Elasticity framework?

MORGRIDGE INSTITUTE FOR RESEARCH
CENTER FOR HIGH THROUGHPUT COMPUTING
WISCONSIN INSTITUTE FOR DISCOVERY

Supercomputers are a significant (and growing) source of computing (processing and storage) resources

Supercomputers are “different”!

- Authentication – Two Factor
- Provisioning – Batch Scheduling
- Worker Network Connectivity – limited if at all
- Allocation – Mapping and Management
- Shared file system – Interference

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

Replace TCP/IP with file transfers for distributed HTC control.

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

Use “streams” of files to implement the control channel between the submission site and the remote execution site

- Need to preserve order

Payloads (input and output sand boxes) are moved separately

- Need to coordinate with control channels

We need integrated Storage Elasticity (Network and I/O bandwidth will come next)

Storage Space is the Key

- Storage for data caching
- Storage for data placement
- Storage for check pointing
- Storage for input sand boxes
- Storage for output sand boxes

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY

High Throughput Computing requires automation as it is a 24-7-365 activity that involves large numbers of jobs

*FLOPY ≠ (60*60*24*7*52)*FLOPS*

300K H*1 J ≠ 1 H*300K J

MORGRIDGE INSTITUTE FOR RESEARCH

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN INSTITUTE FOR DISCOVERY