Integrating HPC and HTC at BNL – A Year Later

Tony Wong

BNL

70 YEARS OF DISCOVERY

A CENTURY OF SERVICE

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Background

- Scientific Data & Computing Center (SDCC) formed in 2016 to leverage existing HTC expertise in the RACF to kickstart support for HPC activities
- Acquired Institutional Cluster (IC) and KNL-based cluster for HPC-based projects
- IC and KNL in stable production configuration
- Available to approved users

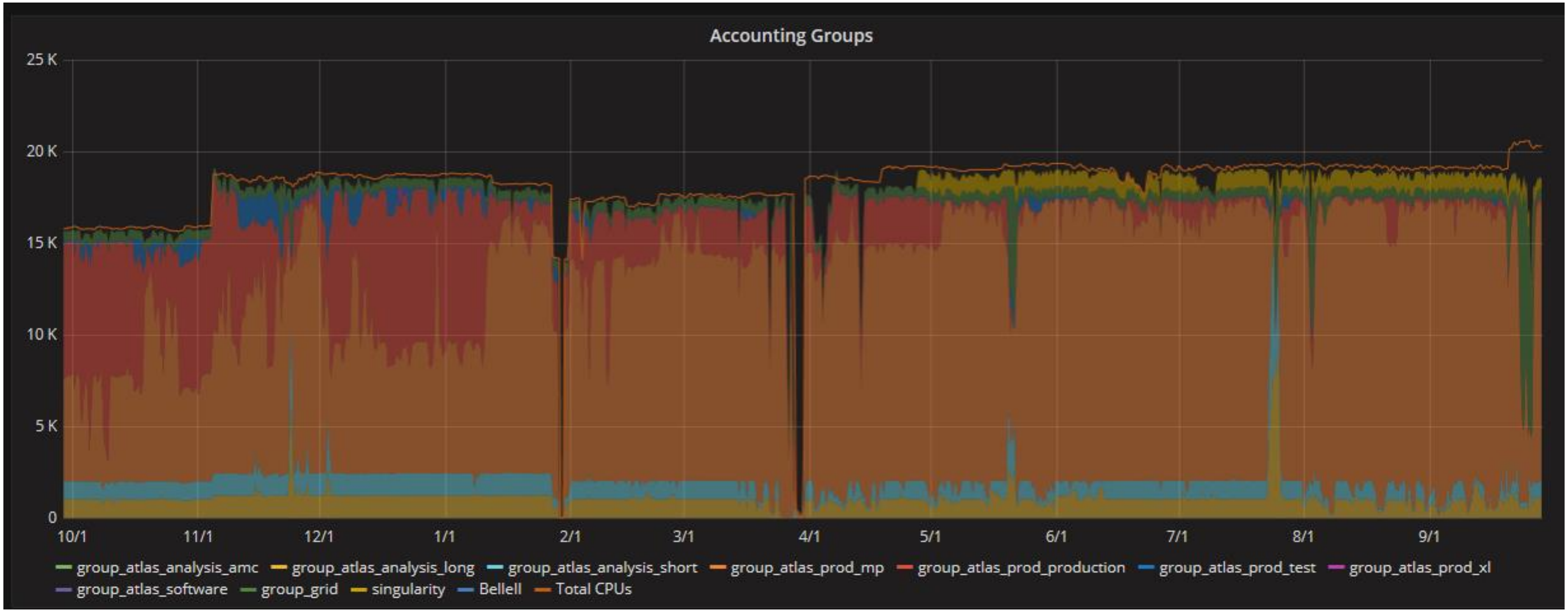
Expanded responsibilities…

- Traditional areas (HTC)
    - HEP and NP
        - RHIC/eRHIC
        - ATLAS
        - Belle-2
    - Intensity Frontier
        - DUNE
        - Daya Bay
    - Cosmic Frontier
        - LSST
        - Legacy projects
- New areas (HPC)
    - LQCD
    - Photon Science
    - Others

…but limited resources

- Manpower shortage
  - Hired one person in April
  - Multiple openings still unfilled
- Resource shortage
  - Current resources fully utilized
  - RHIC experiments falling behind in processing campaigns
  - Encouraged to seek resources on shared clusters at BNL and elsewhere

Fully Utilized ATLAS Farm



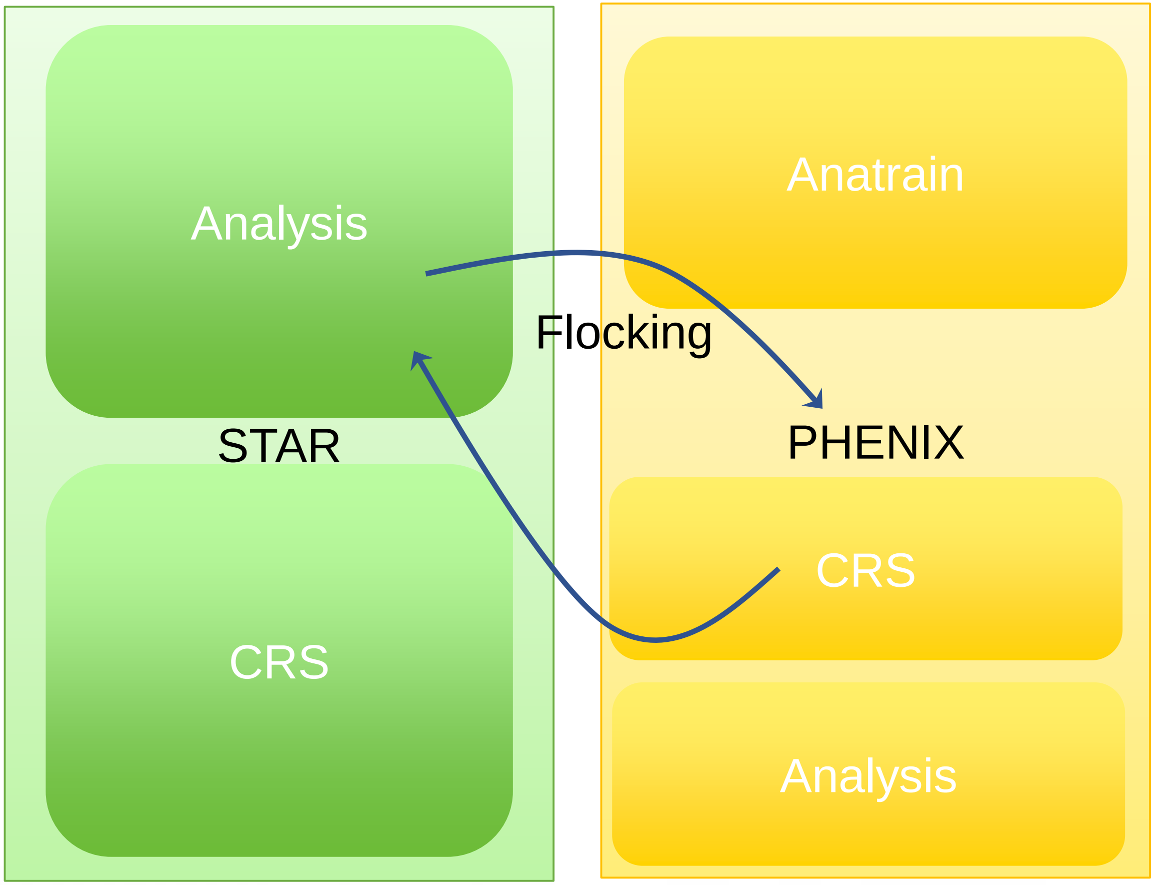
Existing Resources

- Dedicated
  - Custom workloads whose rigid constraints make it difficult for others to use productively
  - Legacy RHIC/ATLAS clusters
- Shared
  - HPC clusters (IC, KNL and others)
  - Recently purchased RHIC/ATLAS resources
  - New general-purpose cluster in 2018

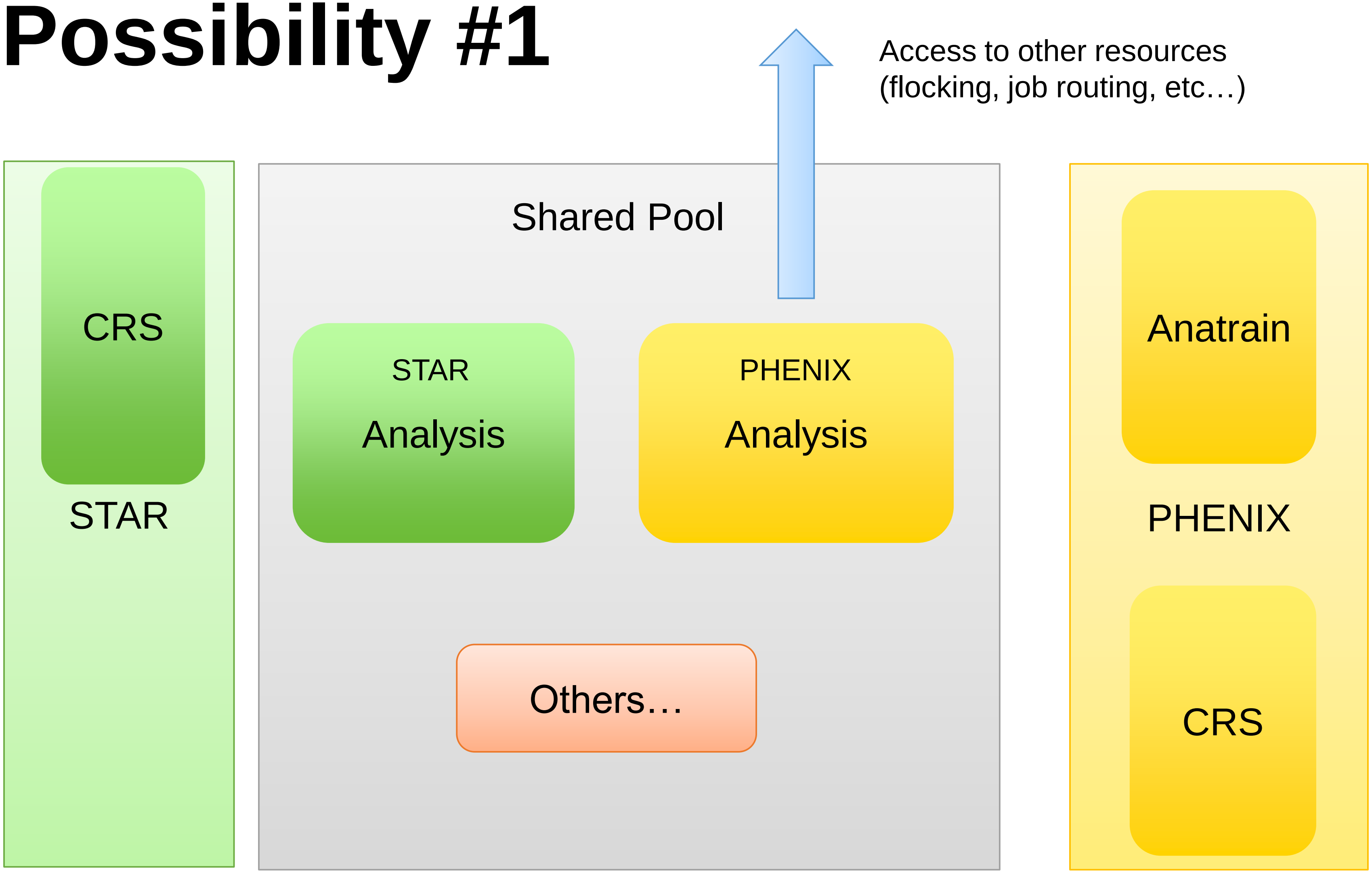
Enabling Resource Sharing

- Integration of cyber infrastructure
  - Discussions on single sign-on to integrate distinct user bases
  - Cross-mounting of disk storage instances
  - Plan to offer access to tape storage via BNLBox
- Rethink HTCondor policy to increase productivity of RHIC/ATLAS clusters. Possibilities are:
  - Collapse multiple HTCondor pools into a single pool and expand usage of hierarchical group quota model deployed on ATLAS Tier 1
  - Increase flocking among multiple Condor pools in existing model
- HTC workloads on HPC clusters
  - Direct access for HPC-adapted workloads
  - Mechanism to submit HTCondor jobs to Slurm at BNL

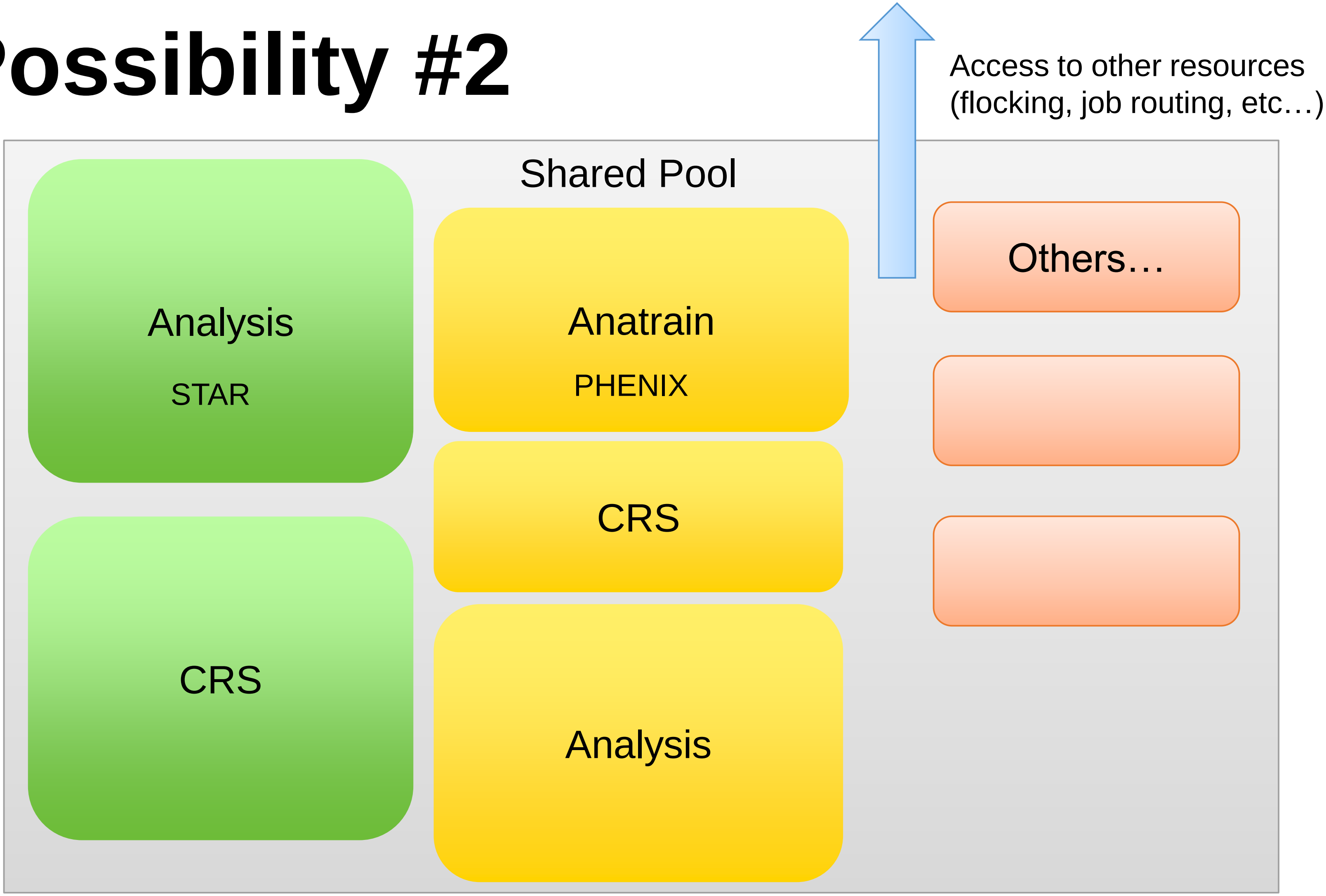
Current HTCondor configuration



Possibility #1



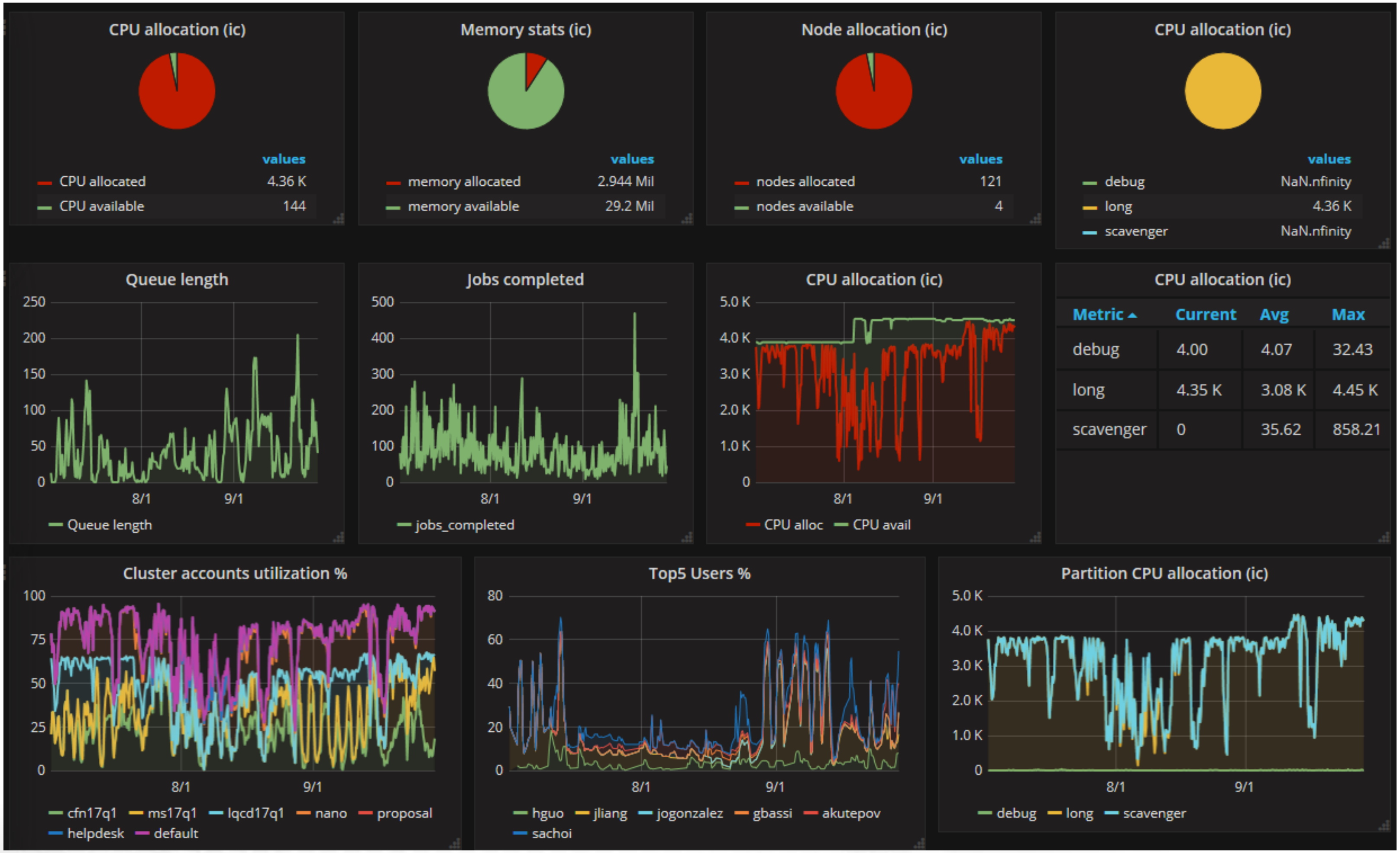
Possibility #2



Institutional Cluster (IC)

- In production since January 2017
- Original cluster with 108 nodes
  - Two Xeon E5-2695v4 (Broadwell) cpu's (36 physical cores)
  - Two Nvidia K80 gpu's
  - 256 GB RAM and ~2 TB SAS disk drive
  - Non-blocking Infiniband EDR fabric
  - 1 PB of GPFS storage with up to 24 GB/s bandwidth via EDR
- Expansion underway
  - Nvidia P100 instead of K80 gpu's
  - First batch of 18 nodes in production since September
  - Another 36 machines purchased in October
  - Full expansion by Spring 2018
- Available to HPC and HTC users

IC Usage



CPU allocation (ic)

	values
🟥 CPU allocated	4.36 K
🟩 CPU available	144

Memory stats (ic)

	values
🟥 memory allocated	2.944 Mil
🟩 memory available	29.2 Mil

Node allocation (ic)

	values
🟥 nodes allocated	121
🟩 nodes available	4

CPU allocation (ic)

	values
🟩 debug	NaN.nfinity
🟨 long	4.36 K
🟦 scavenger	NaN.nfinity

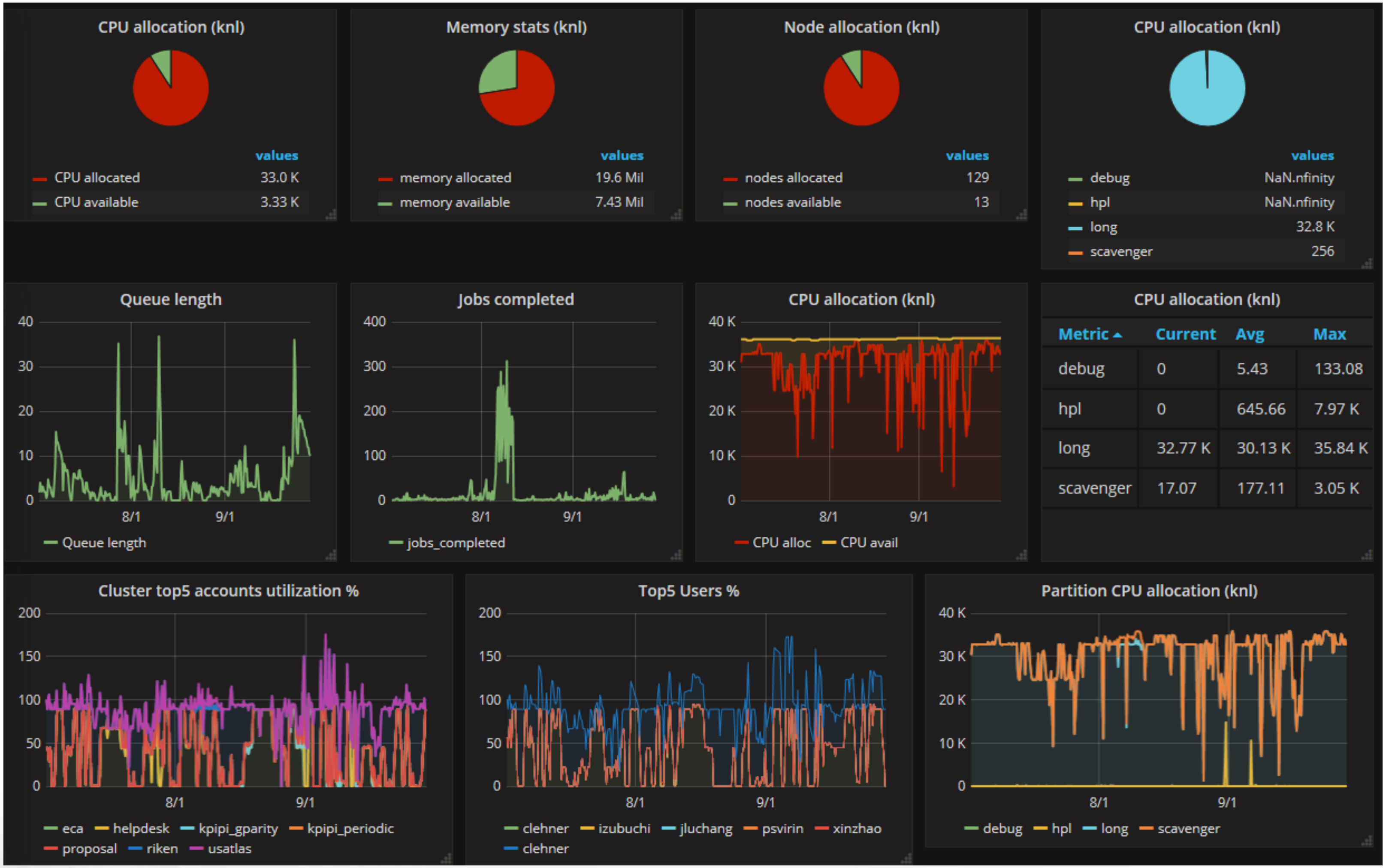
CPU allocation (ic)

Metric ▴	Current	Avg	Max
debug	4.00	4.07	32.43
long	4.35 K	3.08 K	4.45 K
scavenger	0	35.62	858.21

KNL Cluster

- Entered production in June 2017
- 144 nodes
  - One Xeon Phi 7230 cpu (1.3 GHz) with 64 physical cores and 16 GB RAM on chip
  - 2 x 512 GB high-performance SSD drives and 192 GB RAM
  - Dual-rail Intel Omni-Path interconnect fabric with 400 Gbps (nominal) peak aggregate, bi-directional bandwidth
  - Access to IC GPFS storage via custom gateway server and available to users via NFS
- Cluster in useful state, but not optimized
  - Optimization delayed for the sake of stability and availability
  - KNL heavily used by LQCD community
  - Used by ATLAS on an opportunistic basis

KNL Cluster Usage



CPU allocation (knl)

	values
CPU allocated	33.0 K
CPU available	3.33 K

Memory stats (knl)

	values
memory allocated	19.6 Mil
memory available	7.43 Mil

Node allocation (knl)

	values
nodes allocated	129
nodes available	13

CPU allocation (knl)

	values
debug	NaN.nfinity
hpl	NaN.nfinity
long	32.8 K
scavenger	256

Queue length

Jobs completed

CPU allocation (knl)

CPU allocation (knl)

Metric ▴	Current	Avg	Max
debug	0	5.43	133.08
hpl	0	645.66	7.97 K
long	32.77 K	30.13 K	35.84 K
scavenger	17.07	177.11	3.05 K

Cluster top5 accounts utilization %

Top5 Users %

Partition CPU allocation (knl)

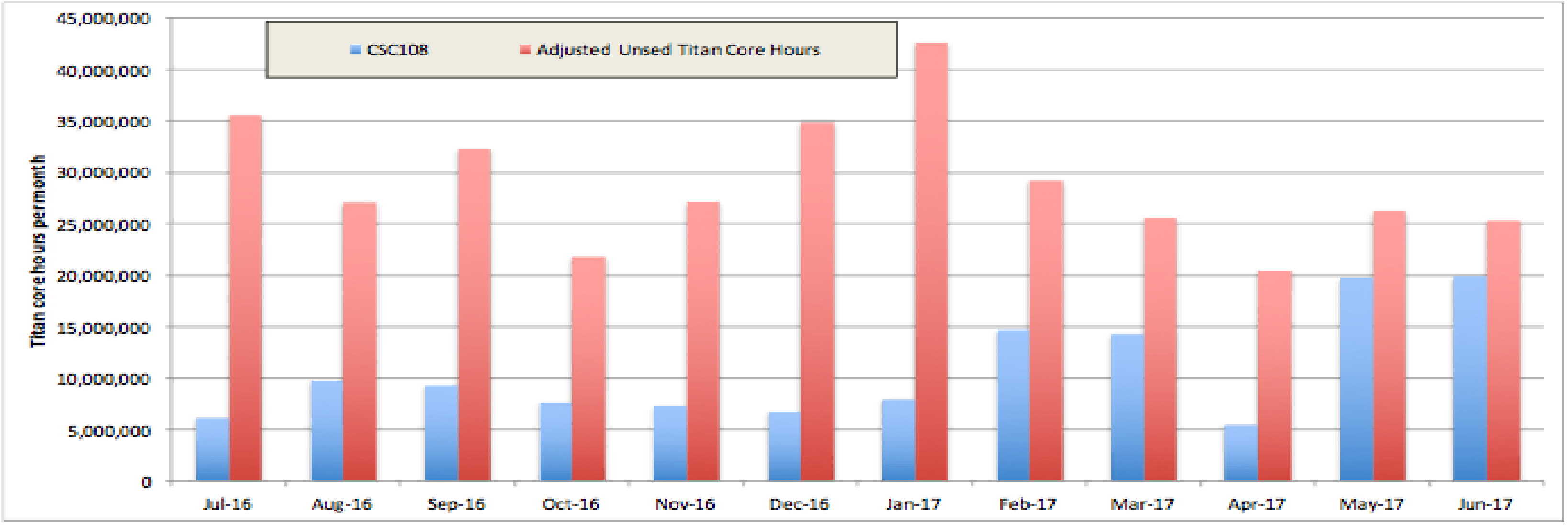
Titan @ ORNL

Slide kindly provided by Sergey Panitkin (BNL)

- 18,688 compute nodes (299,008 logical cores) with GPU's (Cray Xk-7)
- AMD Opteron 6200 @ 2 GHz
- 32 GB RAM per node
- Nvidia K20x
- 32 PB Luster-based disk storage (1 TB/s aggregate throughput)
- 29 PB HPSS-based tape storage
- 27 Pflops peak theoretical performance

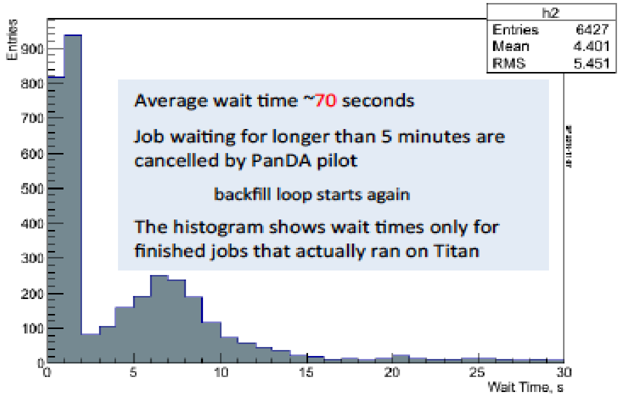
ATLAS on Titan

Slide kindly provided by Sergey Panitkin (BNL)



- Job sizes shaped to backfill opportunistically via PanDA
- Used 129M core-hours from July 2016 to June 2017
- ~2.5% of total available time on Titan
- ~10% of all US-ATLAS computing

Wait time on Titan for ATLAS simulation jobs. Zoom at the first 30 seconds.



Average wait time ~70 seconds

Job waiting for longer than 5 minutes are cancelled by PanDA pilot

backfill loop starts again

The histogram shows wait times only for finished jobs that actually ran on Titan

What’s Next?

- Plans to buy another cluster to be shared between HPC and HTC
  - Likely based on Skylake for ATLAS Tier-1 at BNL
    - Standard dual-socket worker node configuration
    - Add IB EDR interconnect fabric for HPC requirements
  - Available to users in early 2018
- Implement HTCondor changes to increase current RHIC/ATLAS cluster productivity
- Continue to facilitate usage of non-traditional clusters
  - Increase HTC access to HPC resources
  - Employ HPC clusters as jumping point to Leadership Class Facilities

Likely Future Direction

- Broad effort to encourage use of LCF’s to meet computing needs
  - ALCF
  - NERSC
  - ORNL
- Alternative (and possibly) complementary solutions
  - Commercial providers (still in touch with Amazon and Google)
  - Academic clouds